INTEGRATING AI IN A HEALTHCARE WEB APPLICATION

A Project Report
Submitted in partial fulfillment of
the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

by

DONA JACOB	B170280EC
C ARCHANA	B170607EC
ANUPAMA SURESHKUMAR PILLAI	B170038EC
FATHIMATHUSSAHANA	B140137EC

Under the guidance of
Dr Ameer PM

Department of Electronics and Communication Engineering
National Institute of Technology Calicut
Calicut, Kerala, India – 673 601

2020-2021

Dedication

In recent times, the Health Industry around the world has come under immense pressure. With increasing patients, the need to diagnose and give proper treatments is a must for each and every patient.This, unfortunately, does not take into account the fatigue and exhaustion faced by the health worker. Therefore, we dedicate our project to the front line workers by automating some of the simple tasks such diagnosis, prognosis and treatment options.

Acknowledgments

The successful completion of this project depends not only on the efforts of the team but also on the cooperation, encouragement and guidance of a number of people.
First and foremost, we wish to express our sincere appreciation to our Project Guide Dr. Ameer P.M, Asst. Professor, Department of ECE, NITC, for his continuous guidance and persistent help throughout this project.
Our sincere thanks to Dr. PP Deepthi, Head of Department of ECE, NITC for being the constant source of motivation.
Our heartfelt gratitude Dr. V. Sakthivel, our Project Coordinator, and also the Committee members for this wonderful opportunity.
We are also grateful to the Electronics and Communication Department for their knowledge they have imparted to us in the past four years.
Most importantly we also thank our parents and friends for their endless encouragement and support. Last but not the least, we would like to praise and thank God for the grace, love and guidance he has showered upon us throughout the completion of the project.

Dona Jacob
C Archana
Anupama Sureshkumar Pillai
Fathimathussahana

May 2021
National Institute of Technology Calicut

Declaration

We hereby declare that except where specific reference is made to the work of others, the contents of this project report are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This project report is our own work and does not contain any outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Dona Jacob (B170280EC)

C Archana (B170607EC)

Anupama Sureshkumar Pillai (B170038EC)

Fathimathussahana (B140137EC)

NIT Calicut
Date: 22.04.2021

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

Certificate

This is to certify that the project report entitled **Integrating AI in a Healthcare Web Application** submitted by **Dona Jacob (B170280EC), C Archana (B170607EC), Anupama Sureshkumar Pillai (B170038EC), Fathimathussahana (B140137EC)** to National Institute of Technology in Electronics and Communication Engineering, is a bonafide record of the project work carried out by them under my supervision and guidance. The content of the project report, in full or parts have not been submitted to any other institute of university for the award of any degree or diploma.

Dr Ameer P M
(Project Guide)
Dept. of Electronics and Communication Engineering
NIT Calicut

Dr Deepthi P P
(Professor and Head of Department)
Dept. of Electronics and Communication Engineering
NIT Calicut

Date: 11.05.2021

(Office seal)

Abstract

Artificial Intelligence increases the ability of healthcare professionals to better understand the day to day patterns and the need of the patients they care for, for better feedback, guidance and accurate diagnosis. It also helps increase the accessibility of better health care to people and bridges the gap between patients and proper healthcare . Our aim is to devise a healthcare application strictly for medical professionals that makes diagnosis and prognosis of medical conditions more accurate and less tedious for healthcare professionals by automating the work they do now. Our application also aims to provide a medical question answering functionality for quick doubt clearance on medicines and diseases. The diagnosis part of the app contains 2 seperate functionalities. Here we have the detection of tumors on MRI scans using the U-Net Deep Convolutional Network and the detection of anomalies in Chest X-Rays using the CheXNet Deep Convolutional Network. The prognosis part of the app, we use machine learning to predict the risk of heart diseases. Lastly, the app we have a medical question answering functionality which is powered by a BERT model.

Contents

List of Figures

Chapter 1

Introduction

1.1 Literature Review

1.1.1 On detection of tumors on MRIs

Medical image segmentation for detection of brain tumors from magnetic resonance images has been a very significant invention as it is important for detecting and providing the right treatment at the right time. There have been many techniques proposed for classification of tumors such as sliding-window convolutional networks, support vector machine[3] and artificial neural network.
We have chosen the UNET architecture based on the work done by Olaf Ronneberger, Philipp Fischer, and Thomas Brox Computer Science Department and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany.[4] The Authors have found that the network trained end-to-end from very few images outperforms the previous best method i.e a sliding-window convolutional network. The network is also fast as stated by the authors. The segmentation of a 512 x 512 image takes less than a second on a latest GPU.

1.1.2 On detection of anomalies on Chest X Ray

Previous work on Deep learning Architectures in this field

Deep learning based methods are widely used in various sectors and fields today. They have made significant milestones in the biomedical image detection field especially for detection of numerous diseases as stated by Dinggang Shen.
The GoogLeNet and AlexNet neural networks had used the concept of data

augmentation and got an AUC of 0.94. The work was followed by Rajpurkar et al using CheXNet architecture, a deep CNN of 121 layer, to detect 14 different pathologies, including pneumonia in Chest X-Rays. A localisation approach along with feature extraction was used to identify 14 anomalies.
We have implemented this particular architecture [5] in python language and integrated it into this application.

Previous work on CXR dataset

The CXR 8 dataset is provided by the National Institute of Health that comprises 108,948 frontal view X-ray images of 32,17 unique patients. Each of the images have text mined eight disease image labels from the associated radiological reports using NLP. A better alternative in the CXR 14 dataset is used, as this is a more enhanced version provided with an additional 6 categories and images as opposed to the CXR 8 dataset. It also majorly comprises frontal chest x-rays. In conclusion, this dataset is a better representation of a real patient with realistic clinical diagnosis compared to any of the previous Chest X-Rays.

1.1.3 On Prediction of risk of heart disease

In recent years, many achievements have been made in the study of Cardiovascular Disease risk prediction models. In a study conducted in eastern China on cardiovascular risk prediction, they have compared many models including multivariate regression model, CART, Naive Bayes, Random Forest, etc. Keeping the multivariate regression model as a benchmark (AUC = 0.714) for evaluation all other models were compared and random forest gave the highest AUC score of 0.787.[6]

1.1.4 On Medical Question Answering

Previous work in Transfer learning in the field of NLP

Transfer learning, that is, pre-training a neural network model on a task that is already known, for example ImageNet, and then fine-tuning it using the trained neural network as the basis of a new purpose-specific model.
Researches in this field have shown that this technique is useful in so many other fields pertaining to deep learning as well, especially Natural language processing.
The ELMo paper describes a different approach, widely popular in the field of NLP is feature based training.As the name suggests, the trained neural

network outputs word embeddings that is accounted as “features” in the NLP model.[7]

On SQuAD

The Stanford Question Answering Dataset (SQuAD) by Pranav Rajpurkar and Jian Zhang and Konstantin Lopyrev and Percy Liang [8] is a reading comprehension dataset consisting of 100,000+ questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage.

1.2 Motivation

Each year over 12 million adults in the US receive a misdiagnosis, which accounts to 5 percent of adults. Overall rates in the whole world are unknown but research says it accounts to 12-15 percent. The National Academy of Medicine had stated “most people will experience at least one diagnostic error in their lifetime, sometimes with devastating consequences”[9] in the Landmart 2015 report Improving Diagnosis in Healthcare.
Incomplete medical histories, overcrowded health centres, may lead to time and money loss resulting in misdiagnosis or worse ailments. Immune to this, we use Artificial Intelligence to diagnose and predict diseases at a faster rate assisting physicians to make faultless diagnosis.
Similar to how doctors learn through medical school, residency, fellowship and learning from mistakes, AI algorithms also are trained to do arduous tasks. Generally, they perform great for automating the tasks they are trained in and can sometimes also outperform humans. Here, data is first fed with each data having a label for the algorithm, after which the algorithm is exposed to sets of data and corresponding labels which increases the accuracy. [10]

Chapter 2

Problem Definition

2.1 Problem Statement

2.1.1 Statement

Our final objective is to develop a Web Application that automates certain aspects of the healthcare industry :
The entire project is broken into three portions:

1. DIAGNOSIS
    - Detection of tumors on MRI
    - Detection of anomalies on Chest X-Ray
2. PROGNOSIS
    - Risk prediction of cardiovascular disease in 10 years
3. MEDICAL QUESTION ANSWERING
    - A quick doubt clearance functionality about common diseases and treatments from a set of medical documents

2.1.2 App structure

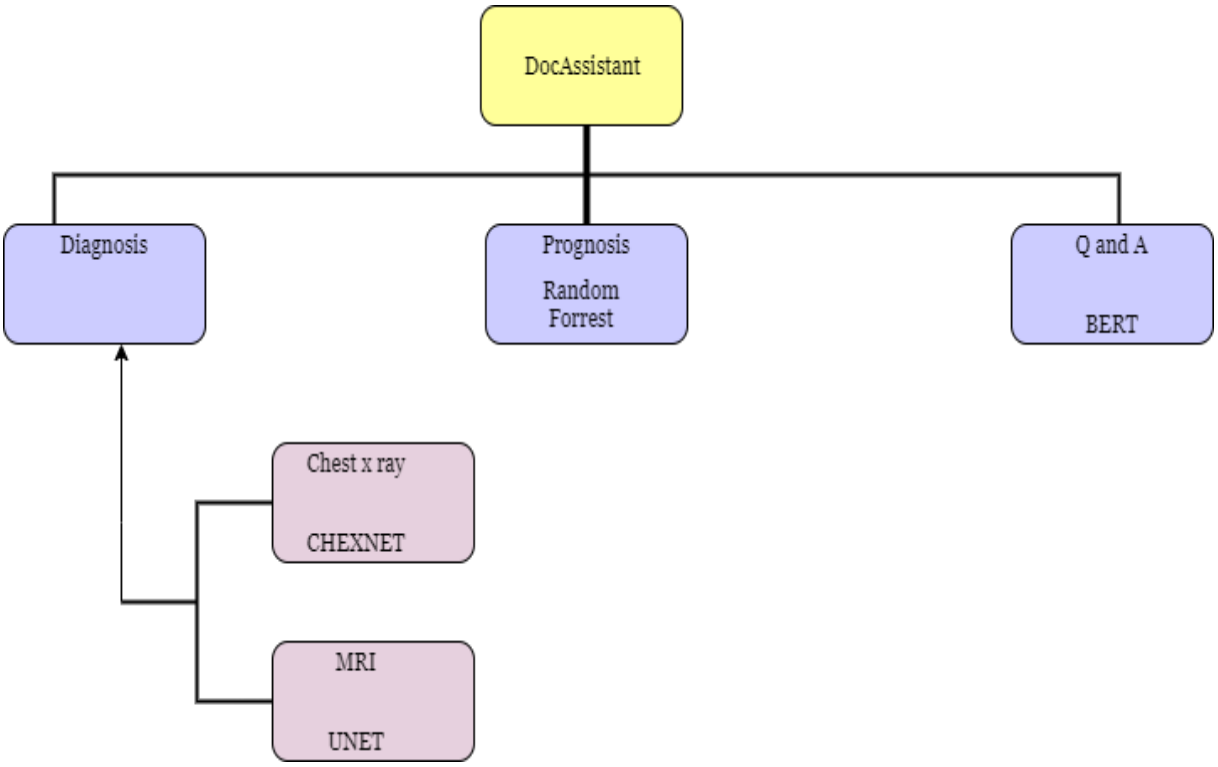

Figure 2.1: Flowchart

Chapter 3

Detection of tumors from MRI using U-Net

3.1 Introduction

Unlike X-rays and CT scans, Magnetic resonance imaging do not use radiation and therefore is considered a noninvasive procedure. Detecting tumors manually by a radiologist is tedious and error prone. Here we introduce a multi-class segmentation model that is dependable and aids in identifying 3 different abnormalities in each image : edema, enhancing and non enhancing tumours

3.2 Dataset Analysis

The Brain Tumor Segmentation Challenge (BraTS) provides the largest fully annotated and publicly available database for model development. The BraTS 2020 dataset comprises 359 training and 125 validation cases.

All BraTS multimodal scans are available as NIfTI files (.nii.gz) and describe FLAIR, T1w, T2w, T1gd which are sequences that correspond to a certain technique of exciting magnetic spins in human body each having different gray scale contrast because certain abnormalities are seen easily in a particular sequence. Refer Fig 3.1.

Training sample used here is split into a first file containing containing 4D array of our MRI with shape (240, 240, 155) the first 3 dimension are 3-D volume,refer fig 3.2,and the 4th dimension is the values for 4 different sequences. The next file in each training case is a label file enclosing a 3D array (with shape 240 X 240 X 155).

3.3 Image segmentation and U-Net

3.3.1 Image Segmentation

The major aim of image segmentation is to divide an image into many image segments on the basis of different properties. It is considered as very important in technology assisted diagnosis systems due to its wide range of applications.

The segmentation is based on similar attributes such as intensity, depth, color or texture resulting in a cluster of regions identifying each label.[11]

Bringing image segmentation into brain MRI analysis leads to the classification of the given data into tumors using inspects with differing intensity.Currently the use of sequences in total four named native T1-weighted (T1), T2-weighted (T2), Gd-enhanced T1-weighted (T1Gd), and Fluid-Attenuated Inversion Recovery (FLAIR) are implicated.The reasoning for using more than one sequence is because each sequence leads to an increased visibility of certain tumor region which may have been compromised with the other sequences. Hence, leading to a more accurate tool for distinguishing.[12]

3.3.2 U-Net

Automatic medical image segmentation are considered the most viable options when it comes to Convolutional Neural Networks with medical diagnosis.This bring us to U-Net that majorly adopts the procedure of labeling our input image, which can be called as image classification.The algorithm recognises the region of interest and renders a label to it.For any algorithm to be successful, we require a large set of training samples whereas in the case of U-Net the inclination is high towards augmenting the current samples efficiently as opposed to having a large training sample.The crux of the algorithm lies in the contracting and expanding path, with the contracting path focusing on absorbing information and the expanding path on precision concentration.[4]

3.3.3 3D U-Net Architecture

U-NET comprises of an analysis/contracting path on the left side and a synthesis/expansive path on the right side. The classic architecture of a Convolutional neural network is emulated in the analysis path. Refer fig 3.3

The architecture of contracting path is that every layer has two 3X3X3 convolutions ,each of which is followed by Rectified Linear unit,which is further

succeeded by 2X2X2 max pooling layer with strides of value 2 in all three dimensions.

The architecture of expansive path is that every layer has an up-convolution of 2X2X2 in strides of two in all 3 dimensions,which is succeeded by two 3X3X3 convolutions and further followed by ReLu.

In every step of down-sampling the number of feature channels is doubled.An up-sampling of the feature map succeeded by 2X2X2 convolution (which cuts down the number of feature channels by half), a concatenation with the equivalently cropped feature map obtained from the contracting path, and two 3x3x3 convolutions, each succeeded by a ReLU is included in each step of contracting path. [13]

3.4 Metrics

3.4.1 Dice coefficient

Once we are done with our model, we want to evaluate the accuracy of our predictions.A major point to be taken into consideration while evaluating the model is to reduce false negatives as they have costly effects in the case of medical diagnosis. Hence, we chose Dice Coefficient as it penalises the false negatives.

Dice score in a nutshell is the measure of how similar the two contour regions are.So it's essentially the overlap or the intersection of the two regions by the total size of the two images.In order to devise a loss function which is to be made as small as possible(i.e,minimized) we take 1-Dice coefficient as loss function. Refer fig 3.4

3.5 Algorithm

1. Generate sub-volumes of the input multi-modal scan,that is generate patches for MR images randomly of [160,160,16] shape
2. Pick patches with greater than 95% tumor
3. Standardize the images to mean 0,standard deviation 1
4. We create the 3D UNET model architecture and compiled the model with the specified loss function,i.e Dice coefficient
5. We run the model on patches

6. We get the result,that is ,the predicted patch on a particular threshold, compare it with the ground truth and do calibration to get almost the same tumor segment as seen in the ground truth. Refer Fig 3.5

7. So for a particular threshold value we get the best prediction.

8. We sew the patches together and move on to doing the prediction on entire scans with that particular threshold

9. We obtain prediction on entire scan

3.6 Observations and results

To test our model we give an input(containing an image,label and model) to obtain the model prediction over whole image. Given that red colour denotes edema presence ,enhancing tumor is shown in blue and non enhancing tumor in green color.

Using this information, when we compare the ground truth and prediction obtained by our model, the visibility of non enhancing and enhancing tumors can be inferred as high whereas the presence of edema is lower than the ground truth. Refer fig 3.6

3.7 Evaluation of Model

Sensitivity is the fraction of true positives (positive results that are correctly predicted) in a particular test. For example, the fraction of people with tumors rightly diagnosed.

Specificity is the fraction of true negatives (negative results that are correctly predicted as negative) e.g., the rate of healthy people who are rightly diagnosed with no disease).

The sensitivity and specificity of our prediction is shown below. Refer fig 3.7.

3.8 Inference

From the specificity and sensitivity values obtained we notice that specificity of all three conditions i.e, edema, enhancing and non enhancing tumors are

high hence we can infer that the probability of the model giving a positive result for a healthy patient is unlikely.

If the model gives a positive result in a particular test case it means the probability of having that particular condition is high

The sensitivity of enhancing tumor and non enhancing tumor is quite high which means that when our model gives a negative result, it means the result is in fact negative.

For the case of edema it is low (which was evident from the figure as the prediction had only faint red lines contrary to the ground truth).In conclusion our model gives a fairly good prediction with high accuracy in detecting tumors.

3.9 Figures

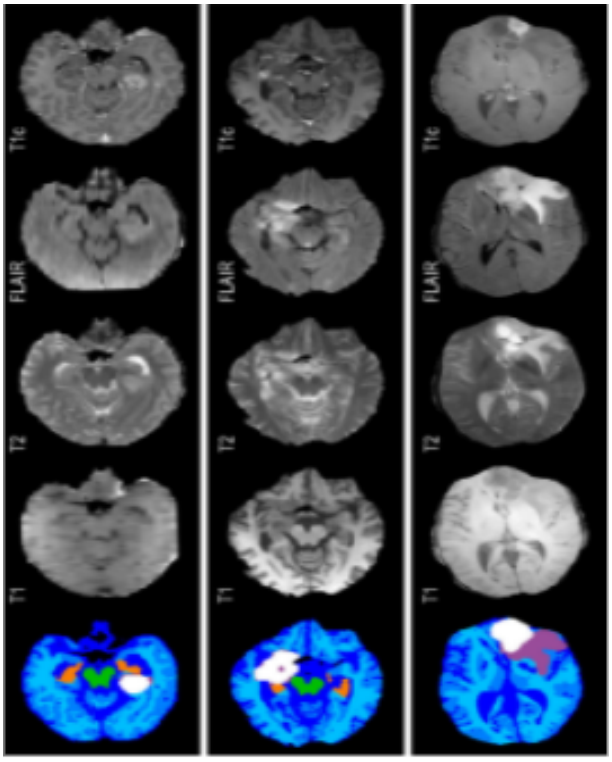

Figure 3.1: NIfTI files (.nii.gz) a)T1 b)T2 c)FLAIR d)T1c

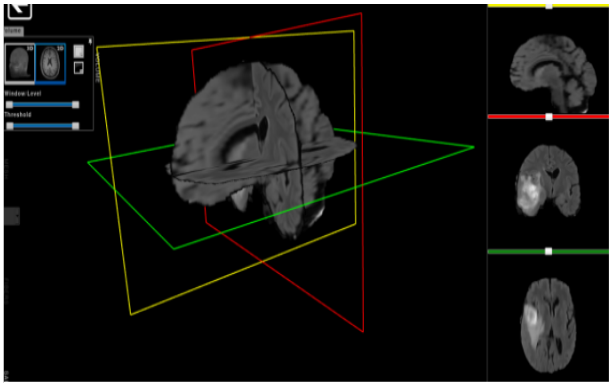

Figure 3.2: BRATS_001 (.nii)

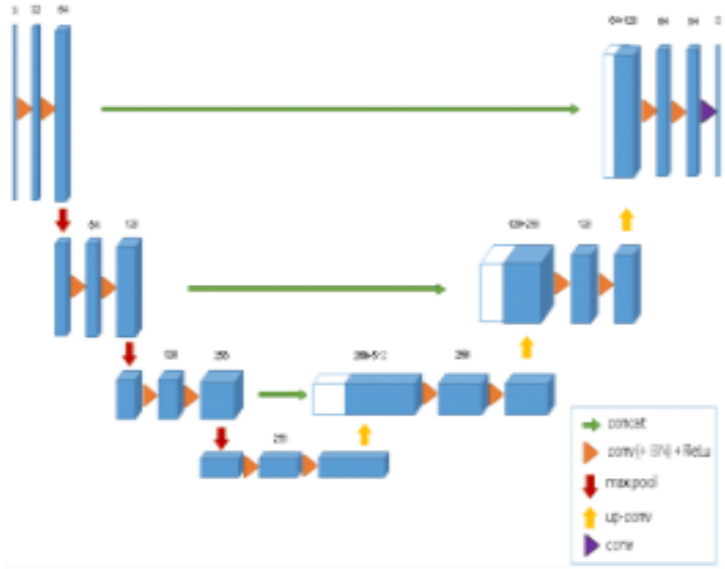

Figure 3.3: 3D UNET architecture

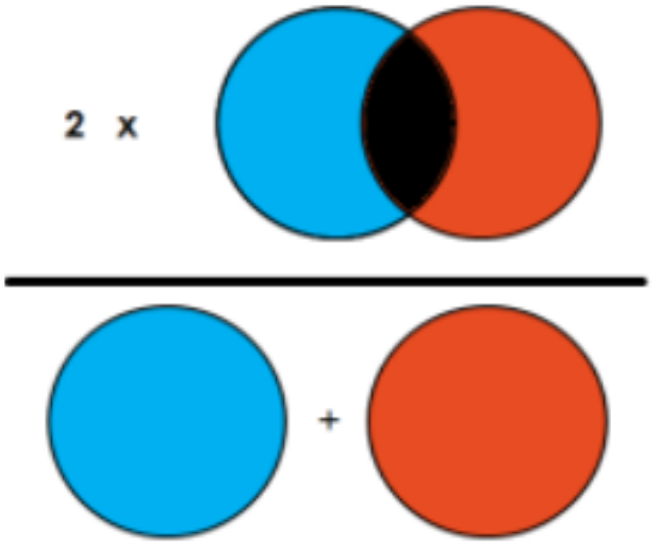

Figure 3.4: Dice coefficient illustration

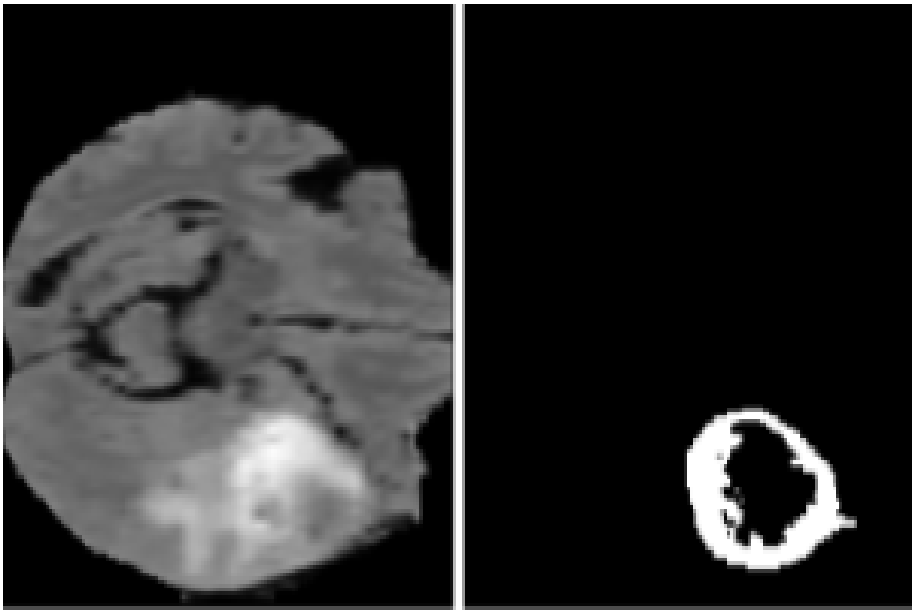

Figure 3.5: Subvolume generated and corresponding tumor patch

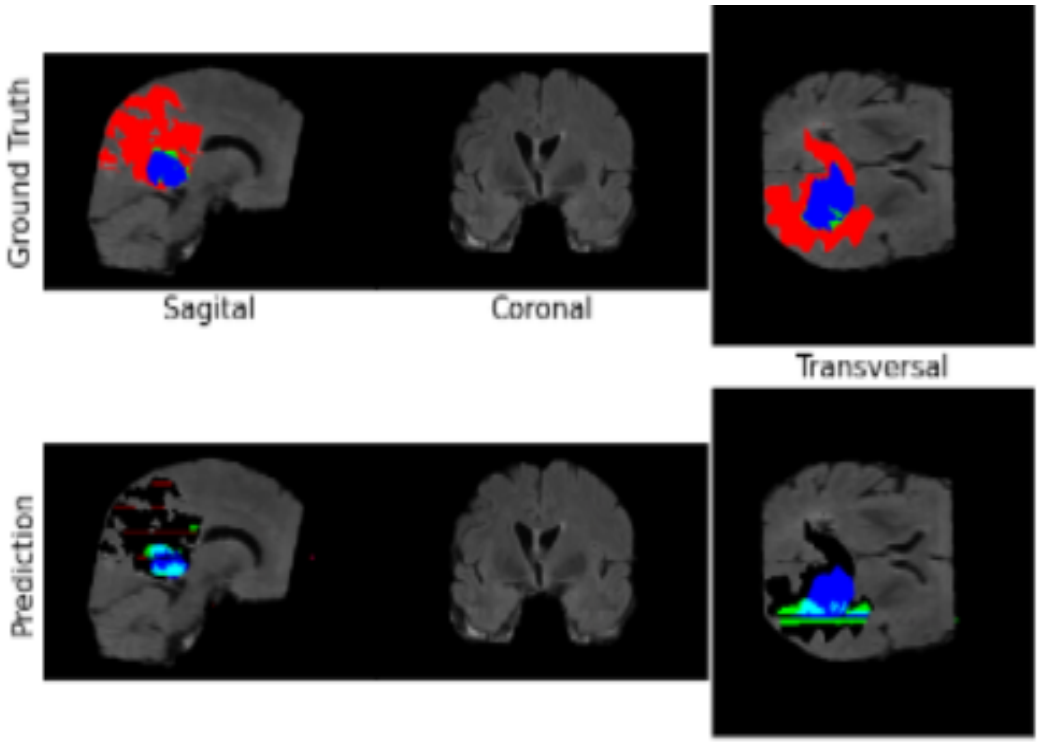

Figure 3.6: Ground truth and prediction

	Edema	Non-Enhancing Tumor	Enhancing Tumor
Sensitivity	0.0394	0.524	0.8277
Specificity	0.9994	0.9859	0.9924

Figure 3.7: Evaluation

Chapter 4

Detection of Anomalies from Chest X-Rays

4.1 Introduction

The second objective of our major project is the detection of anomalies from chest X-Rays. Accurate detection of anomalies present in chest X-Rays is vital for disease detection. Manual detection of these conditions by a radiologist could be tiresome and error-prone. Here we introduce a programmable aid which can detect the presence of 14 conditions:Atelectasis, effusion, cardiomegaly, pneumonia, pleural thickening,pneumothorax, infiltration, mass, nodule, Fibrosis, Consolidation, Hernia, Edema, Emphysema.

4.2 Dataset Analysis

The National Institutes of Health Chest X-Ray Dataset[14] is a collection of large number of X-Ray images with disease labels from patients. These labels are created by the authors via the process of text mining in radiology reports with Natural language processing technique with an expected accuracy of greater than 90 percent in labelling and it is considered suitable for weakly supervised learning. In this project, we have used 15,000 X-Ray images and corresponding labels from the NIH Chest X-ray dataset. The labels are of 14 pathological conditions: Emphysema, cardiomegaly, effusion, hernia, mass, pneumothorax, pneumonia, pleural thickening, fibrosis, edema, consolidation, atelectasis and nodule.[14]

4.3 Pre-processing

Since we are using a pre-trained model from PyTorch as a part of our transfer learning procedure, we have to preprocess our dataset in such a way that it is compatible with the previous model.

First, we resize and centre crop the image to 224 X 224 size, convert it to a tensor and then normalize it. The resulting images we obtain after passing them through these transforms are tensors that can be input into our network. The training transformations are similar but with the addition of random augmentations.

4.4 Transfer learning

In transfer learning, we obtain a model trained on larger datasets and transfer its knowledge to a smaller dataset. Here the convolutional layers extract basic features such as edges, contours, patterns and the layers that come after that recognise the finer details. There would always be low-level features in common and therefore taking information from models trained on similar but larger datasets and transferring their features to the current task is expected to give better results.

In our project, we have used the PyTorch library for transfer learning. Pytorch contains a number of models which are trained on millions of images from ImageNet. Here we load in pre-trained weights from Densenet-121 network trained on ImageNet, freeze all weights in lower convolutional layers. Then the rest of the custom layers are then trained by which we optimize the model.

4.5 CheXNet Model

In the DenseNet CNN every layer of network is extensively connected with every other layer in a feed forward manner[15]. In terms of Model, the CheXNet model is the same as DenseNet -121 but in terms of Architechture, CheXNet and Densenet are same except for the output layer. The output layer of CheXNet consists of a softmax layer containing 14 categories of different lung diseases. The algorithm of Transfer learning is used on CheXNet. The Densenet - 121 model is pre trained on ImageNet dataset, which contains about 20,000 images. The information obtained from this model is transferred to the CheXNet model which is trained on Chest X-ray 14 dataset.[16]

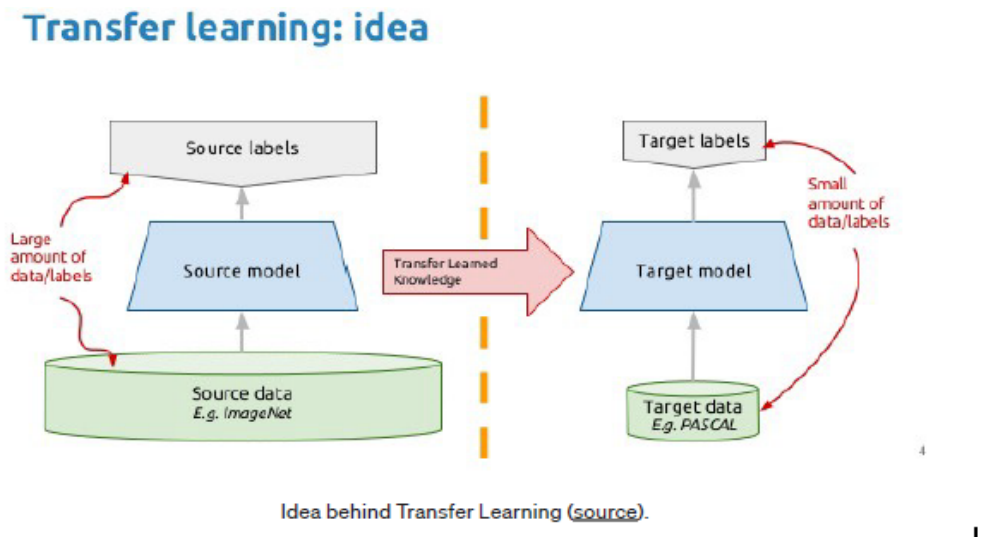

Figure 4.1: Transfer Learning Concept[1]

4.6 Algorithm

- Pre-trained torchvision model, that is densenet 121 pretrained on ImageNet is obtained.
- The custom layers of this densenet model (the last layers) are modified to include an output layer with softmax.
- The next step is Image preprocessing where the following is done:
    - PyTorch uses transforms to preprocess an image
    - It first resizes an image into size 224 X 224 (image size of data in imagenet)
    - Normalization
- We further augment the images using the transforms library in Pytorch where we randomly flip our images
- Training the model with CXR-14 dataset
- The model is trained on parameters:
    - Optimizer: Stochastic gradient descent
    - Loss criterion: Binary Cross entropy
    - Batch size: 16

    - Initial learning Rate: 0.01

- The model with lowest validation loss is chosen and weights are checkpointed [17]

4.7 AUC-Evaluation

The performance metrics we have chosen to evaluate our model is AUC scores. It stands for the area under the receiver operating characteristics curve(ROC). It is, in general, the best-suited performance measure for classification tasks and imparts a total measure of performance across classification thresholds. It shows us the capability of the model to distinguish between the classes. The higher the AUC score is, the better the model is at distinguishing between patients with and without a particular pathological condition. The AUC scores can be explained in such a way that:

- Closer it is to 1, it means the model has a good measure of separability, which is the desired characteristic

- Closer it is to 0, it has the worst measure of separability, which is the reciprocated result.

- When AUC is 0.5, there is no class separation interpreted by the model,which makes the model unreliable.

4.8 Observations

The presence of a particular anomaly is displayed as the final output. To understand the predictions, we produce a Heatmap, which enables us to visually analyze the areas of the image which have the indicative part of the disease.

We use a method called Class Activation mapping for convolutional neural networks with global average pooling. Here we teach the CNNs to perform object localization. These CAMs helps us to 'visualize the predicted class scores on any given image, highlighting the discriminative object parts detected by the CNN'. Then we use these activations and convert them into a probability for each of the 7x7 subregions and then calculate ln(p_subregion / p_baseline) for each of the 7x7 subregions, where p_subregion is the probability of disease based on that subregion and p_baseline is the population baseline probability of the disease.

Case 1: We search for Cardiomegaly in a particular test image (00000075_001.png) (Fig 4.2)

- Output: Heatmap showing presence of cardiomegaly
- The probability that the condition exists: 12.7%
- Table (Fig 4.3) showing the predicted probability values of all conditions present in the test case and their ground truths.
- Here, we obtain a 12.7% probability for Cardiomegaly, and that is higher than the baseline probability of the NIH dataset(2.5%) and therefore the test case is positive for Cardiomegaly. (Fig 4.8: table of baseline probabilities)

Case 2: We search for Pneumonia in a test image (00000211_013.png) (Fig 4.4)

- Output: Heatmap showing presence of Pneumonia
- The probability that the condition exists: 5.1%
- Table (Fig 4.5) showing the predicted probability values of all conditions present in the test case and their ground truths
- Here, we obtain a 5.1% probability for Pneumonia, and that is higher than the baseline probability of the NIH dataset(1.2%) and therefore the test case is positive for Pneumonia. (Fig 4.8: table of baseline probabilities)

Case 3: We search for Emphysema in a test image(00000020_000.png)(Fig 4.6)

- Output: Heatmap showing presence of Emphysema
- The probability that the condition existed: 1.2%
- Table (Fig 4.7)showing the predicted probability values of all conditions present in the test case and their ground truths.
- Here, we obtain a 1.2% probability for Emphysema, and that is higher than the baseline probability of the NIH dataset(2.2%) and therefore the test case is positive for Emphysema. (Fig 4.8: table of baseline probabilities)

4.9 Inference

- The heatmap is structured in such a way that yellow/green regions have higher positive values(presence of anomaly) and blue/purple shows negative values.
- The observations of 3 example cases and corresponding results are described. We observe the first test case was positive for cardiomegaly, second test case was positive for pneumonia and third test case was negative for emphysema.
- Predicted probabilities of all 14 conditions are also displayed. We conclude whether the test subject is positive or negative for an anomaly by comparing the predicted probability to baseline probabilities(fig 4.8) of NIH dataset.
- The performance metrics we have chosen to evaluate our model is AUC scores. The individual AUC score of the model predicting each of the conditions are displayed. (Fig 4.9)
- We could notice that for all conditions the AUC scores are in range: 0.6 to 0.9.
- We have 12 conditions with AUC greater than 0.7, which means the model has a good measure of separability, which is the desired characteristic.
- Mass and Nodule conditions have AUC between 0.6 to 0.7 which means the model has an average measure of separability for these conditions.

4.10 Figures

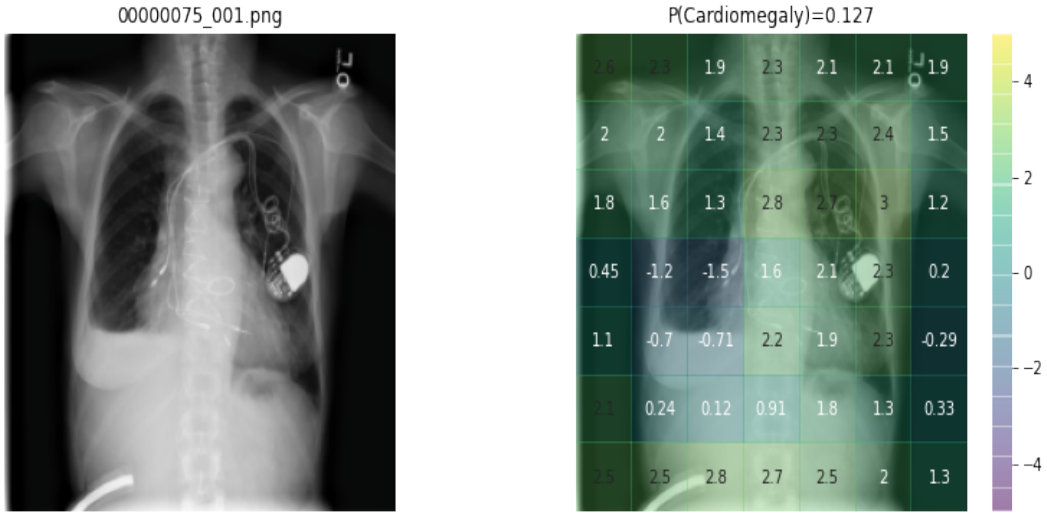

Figure 4.2: Cardiomegaly

Finding	Predicted Probability	Ground Truth
Effusion	0.675	False
Cardiomegaly	0.127	True
Atelectasis	0.109	False
Infiltration	0.102	False
Pleural_Thickening	0.075	False
Pneumothorax	0.071	False
Mass	0.035	False
Consolidation	0.015	False
Emphysema	0.013	False
Nodule	0.008	False
Fibrosis	0.008	False
Pneumonia	0.005	False
Edema	0.002	False
Hernia	0.002	False

Figure 4.3: Predicted Probability (all 14) and Ground Truth(test case :1)

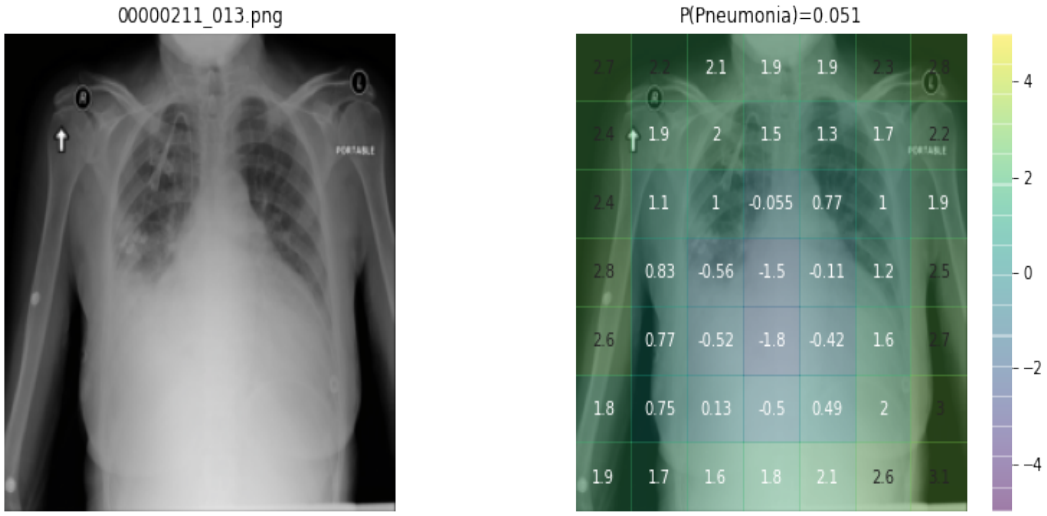

Figure 4.4: Pneumonia

Finding	Predicted Probability	Ground Truth
Infiltration	0.527	True
Effusion	0.473	True
Edema	0.312	True
Consolidation	0.266	False
Atelectasis	0.223	False
Cardiomegaly	0.147	True
Mass	0.059	False
Pneumonia	0.051	True
Pleural_Thickening	0.039	False
Nodule	0.018	False
Fibrosis	0.016	False
Pneumothorax	0.010	False
Emphysema	0.005	False
Hernia	0.002	False

Figure 4.5: Predicted probability(all 14) and Ground truth(test case :2)

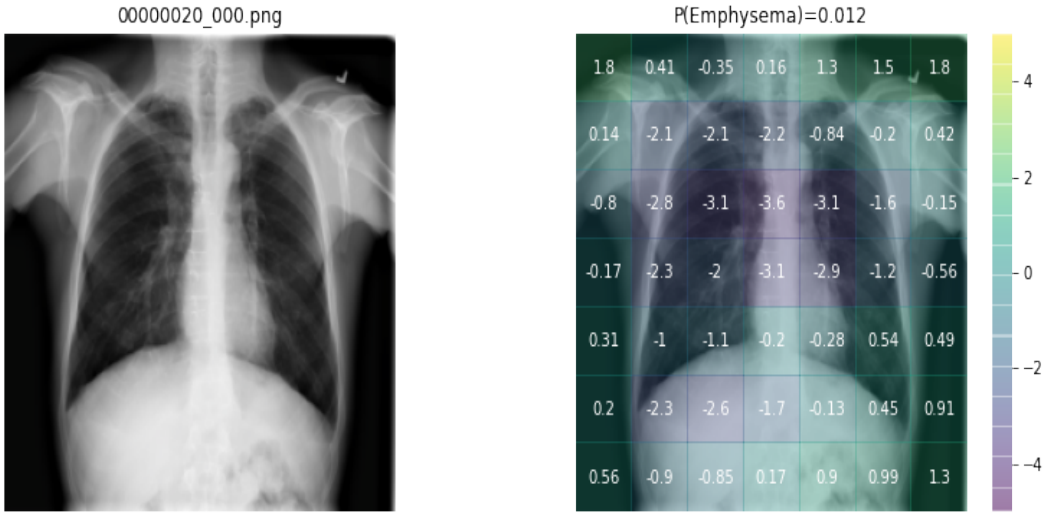

Figure 4.6: Emphysema

Finding	Predicted Probability	Ground Truth
Fibrosis	0.133	False
Infiltration	0.112	False
Pneumothorax	0.078	False
Nodule	0.031	False
Pleural_Thickening	0.030	True
Emphysema	0.012	False
Consolidation	0.010	False
Pneumonia	0.008	False
Atelectasis	0.004	False
Effusion	0.004	False
Mass	0.004	False
Hernia	0.003	False
Cardiomegaly	0.001	False
Edema	0.001	False

Figure 4.7: Predicted probability (all 14) and Ground truth(test case :3)

Characteristic	IU	MSH	NIH
Patient demographics			
No. patient radiographs	3,807	42,396	112,120
No. patients	3,683	12,904	30,805
Age, mean (SD), years	49.6 (17.0)	63.2 (16.5)	46.9 (16.6)
No. females (%)	643 (57.3%)	18,993 (44.8%)	48,780 (43.5%)
Image diagnosis frequencies			
Pneumonia, No. (%)	39 (1.0%)	14,515 (34.2%)	1,353 (1.2%)
Emphysema, No. (%)	62 (1.6%)	1,308 (3.1%)	2,516 (2.2%)
Effusion, No. (%)	142 (3.7%)	19,536 (46.1%)	13,307 (11.9%)
Consolidation, No. (%)	26 (0.7%)	25,318 (59.7%)	4,667 (4.2%)
Nodule, No. (%)	104 (2.7%)	569 (1.3%)	6,323 (5.6%)
Atelectasis, No. (%)	307 (8.1%)	16,713 (39.4%)	11,535 (10.3%)
Edema, No. (%)	45 (1.2%)	7,144 (16.9%)	2,303 (2.1%)
Cardiomegaly, No. (%)	328 (8.6%)	14,285 (33.7%)	2,772 (2.5%)
Hernia, No. (%)	46 (1.2%)	228 (0.5%)	227 (0.2%)

*Sex data available for 1,122 / 3,807 IU, 42,383 / 42,396 MSH; age data available for 112,077 / 112,120 NIH.
Abbreviations: IU, Indiana University Network for Patient Care; MSH, Mount Sinai Hospital; NIH, National Institutes of Health Clinical Center; No., number.

https://doi.org/10.1371/journal.pmed.1002683.t001

Figure 4.8: Baseline Probability of 14 conditions of NIH Dataset [2]

	label	auc
0	Atelectasis	0.786392
1	Cardiomegaly	0.894344
2	Consolidation	0.776439
3	Edema	0.894503
4	Effusion	0.860853
5	Emphysema	0.798212
6	Fibrosis	0.758051
7	Hernia	0.797390
8	Infiltration	0.712198
9	Mass	0.683773
10	Nodule	0.692331
11	Pleural_Thickening	0.785400
12	Pneumonia	0.711823
13	Pneumothorax	0.749336

Figure 4.9: AUC scores

Chapter 5

Risk prediction of Heart Diseases

5.1 Introduction

Here we discuss the third objective of our major project. We create a prognostic model using machine learning to predict the risk of heart disease of patients in the next few years.

A Random Forest model is used to determine the above objective and the model hyperparameters are tuned on the basis of the optimal C-index. The next objective is to figure out which factors in the dataset contribute most to the risk of survival for a given patient. We use the SHAP library which is a method used to help read data by quantifying and visualising each feature of the prediction.

5.2 Dataset description

This dataset comprises data from 4 databases - the Long Beach and Cleveland Clinic Foundation, the Hungarian Institute of Cardiology. Budapest, University Hospital, Zurich and Basel, Switzerland[18]. The presence of heart disease is implied in terms of binary, representing 0 for no disease and 1 for its presence.

The features within the dataset are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrographic results, max heart rate, exercise-induced angina, count of major vessels, thalassemia, ST depression(old peak) and slope of ST segment.

5.3 Exploratory data analysis

A basic data exploration is done, to understand the features of the dataset better.

- Heatmap
  - Seaborn library is used for visualization for the plotting of the heatmap.
  - The plotted heatmap is used to display the correlation between the features. (fig 5.1)
  - As observed in the above figure, we can infer that chest pain and max heart rate achieved are the most correlated features to the target value whereas exercise-induced angina and old peak value are the least correlated features to the target.
- Missing Data
  - Before we move on to the modeling, a major step is to do is organise the data.
  - Hence, we check for missing data. The observation led to no rows missing .
  - Therefore, we can conclude that the dataset contains no missing data and requires no imputation.

5.4 Random forest

Random forests are well known models in machine learning used for both classification as well as regression problems. They are a collection of decision trees, used while training and the output is taken as an average prediction of each tree. Random Forests are preferred over Decision Trees as the latter is more prone to overfitting. Hence, we use Random Forests for a more accurate result.

5.5 Hyperparameter tuning

The default hyperparameters of a random forest model may lead to overfitting, so optimization of 3 major hyper parameters (n_estimators, max_depth and min_sample_leaf) is done using hyperparameter grid search.[19]

For each target parameter, a set of possible values are defined. Next the model is trained, and it is evaluated via C-index for all viable combinations of hyperparameters and the best set is selected. In our case we chose n_estimators: 500, max_depth : 15, min_samples_leaf: 1.

5.6 C-Index

C-index in an evaluation scheme, that estimates the ability of the model to differentiate amongst all the classes. The evaluation is done by accessing all the patients in pairs (A,B), where the model has concluded that Patient A has a higher risk score in comparison to Patient B and the observed data states that Patient A has heart disease and Patient B did not have any heart disease.

5.7 Steps

- The data is imported followed by performing exploratory data analysis on it.
- It is further split into train, test and validation datasets in the ratio 60:20:20
- Next the random forest model for classification is constructed and the hyperparameters are tuned by a grid search function using the C-Index to evaluate each tested model.
- The risk scores of the test subjects are calculated and a binary output 1 for presence of disease and 0 for no disease. The accuracy matrix is calculated and ROC curve is plotted.[20]
- The data is visualised using SHAP which helps in explaining the prediction by quantifying the contribution of each feature to the prediction. The force plot of an individual patient is plotted using SHAP values [21]
- The model output as a whole can also be interpreted using SHAP values.

5.8 Observations

- We predict the target value to be

  - 0 if the risk scores is low, i.e if the patient has no risk on having a heart disease for a period of time.
  - 1 if the risk score is high, i.e if the patient has high survival risk.

- We visualise the force plot of a patient, that is we plot a graph that shows which of the features contribute more to the survival risk of that particular patient.
- On considering the model as a whole, we plot the summary of SHAP values(i.e,the impact on model output) of all features.
- The accuracy matrix is plotted. (Fig 5.2)
- The ROC curve is plotted and AUC is calculated. (Fig 5.3)
- The AUC scores are calculated to be 0.964.

5.9 Results

TEST CASE 1

1. PATIENT ID NO: 617
2. The risk score of test case 1 is calculated to be 0.962667
3. Since the risk score is very high the predicted output of our model is 1.
4. We also print a force plot using SHAP library so as to get an efficient summary of this particular patient's prediction. (Fig 5.4)

- The force plot we plotted gives us the following inferences:

1. The features marked in red contribute more to the survival risk. In this particular case we see that chest_pain_type contributes maximum to the elevated risk score followed by thalassemia, st_depression etc. Increase in any of these features increases the risk score
2. The features marked in blue lead to a lower survival risk. In this case we don't see any helpful features due the patients high risk score.

TEST CASE 2

1. PATIENT ID NO: 842

2. The risk score of test case 2 is calculated to be 0.24220
3. Since the risk score is low the predicted output of our model is 0.
4. We also print a force plot using SHAP library so as to get an efficient summary of this particular patient's prediction. (Fig 5.5)

- The force plot we plotted gives us the following inferences:

1. The features marked in red contributes more to the survival risk.In this particular case we see that chest_pain_type contributes maximum to the elevated risk score followed by max_heart_rate_acheived, exercise_induced_angina, etc. Increase in any of these features increases the risk score (bad for the patient).
2. The features marked in blue leads to a lower survival risk. In this case we can see factors such as st_depression, thalassemia etc whose increase in value contributes to decrease in survival risk scores(good for the patient).

TEST CASE 3

1. PATIENT ID NO: 832
2. The risk score of test case 3 is calculated to be 0.564154
3. The risk score is average but is predicted to give an output 1
4. We also print a force plot using SHAP library so as to get an efficient summary of this particular patient's prediction. (Fig 5.6)

- The force plot we plotted gives us the following inferences:

1. The features marked in red contributes more to the survival risk.In this particular case we see that chest_pain_type contributes maximum to the elevated risk score followed by st_slope, max_heart_rate_acheived, exercise_induced_angina etc. Increase in any of these features increases the risk score (bad for the patient).
2. The features marked in blue leads to a lower survival risk.In this case we can see factors such as, thalassemia, num_major_vessels, etc whose increase in value contributes to decrease in survival risk scores(good for the patient).

5.10 Inference

1. From the Accuracy matrix where the y-axis represents the true label and the x-axis represents the predicted label we can observe that,

   (a) 1.46% of the total cases have been predicted to get heart disease whereas they haven't got the disease in reality. We can even observe that 9.27% of the total cases have been predicted to not get heart disease whereas in reality they did get the disease.

   (b) Hence, a total accuracy of 89.269% is observed.

   In conclusion, we can say that the model gives good predictions with high accuracy in detecting risk of heart disease.

2. ROC curve

   (a) The curve shows the performance of the model with respect to two parameters:

   True Positive Rate and False Positive Rate.

   (b) True Positive rate also known as recall is defined as the total number of true positives as a whole divided by the summation of true positives and false negatives. Hence, True Positive Rate signifies the probability that the positive prediction(1) is the true label(1).

   (c) False Positive rate is defined as the total number of false positives divided by the summation of false positives and true negatives. Hence, it leads to the probability of predicting a positive result(1) when the actual observation is a negative result(0).

   (d) The graph shows the rate of True Positive cases as well as False Positive cases.

   Given that the area under the ROC curve of a test calculates the test's discriminative ability, we infer that the test is more efficient as the ROC curve is more closer to the upper left corner as a result the AUROC will be more, which is the case observed in fig 5.3.

5.11 Figures

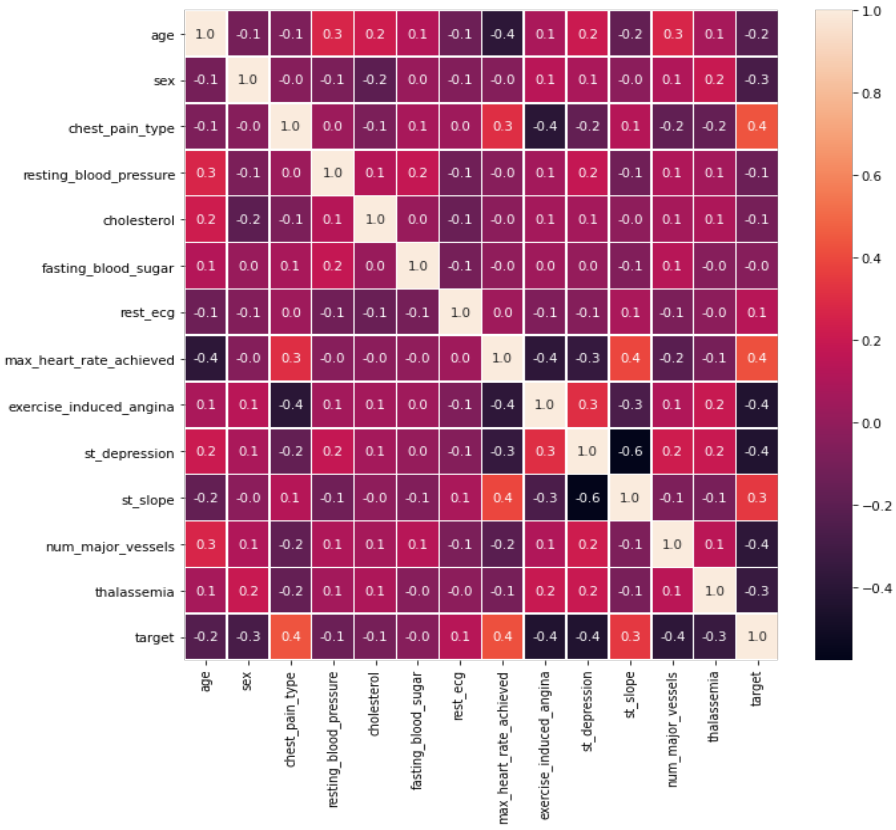

Figure 5.1: HEATMAP

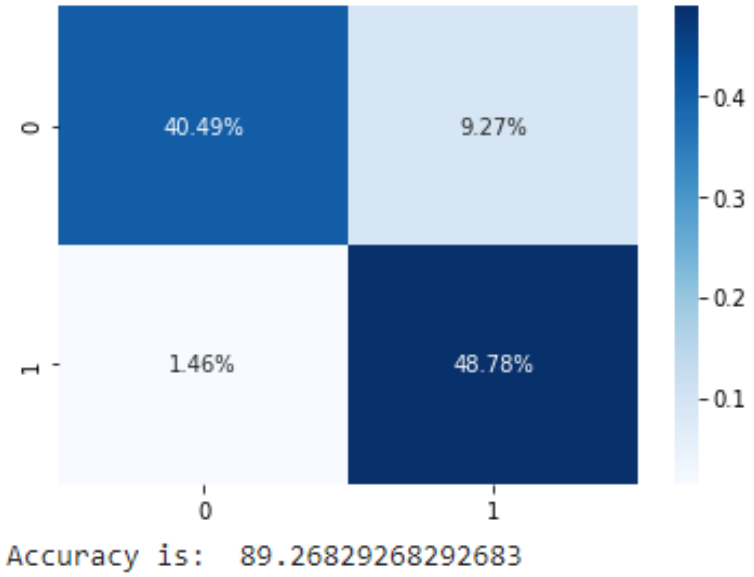

Figure 5.2: ACCURACY

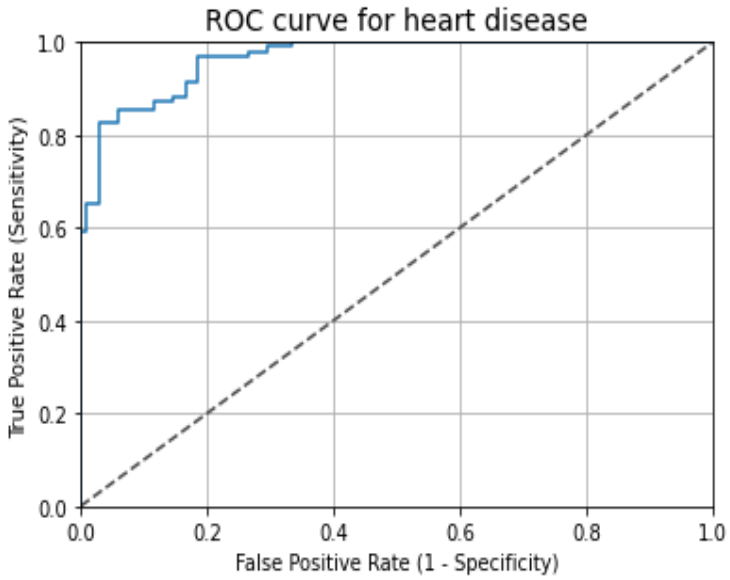

Figure 5.3: ROC CURVE

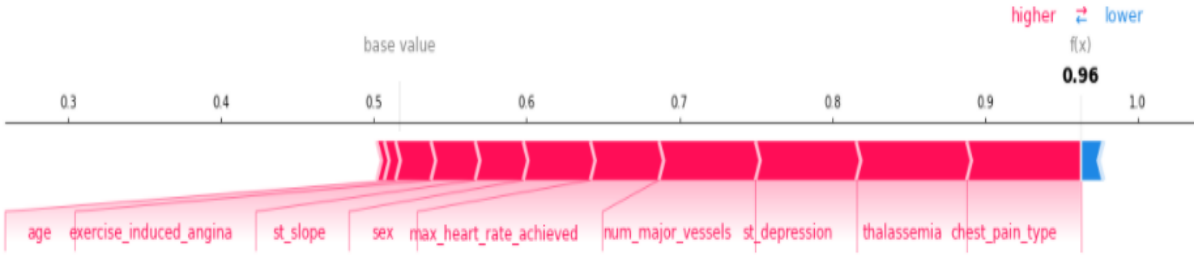

Figure 5.4: Test case :1

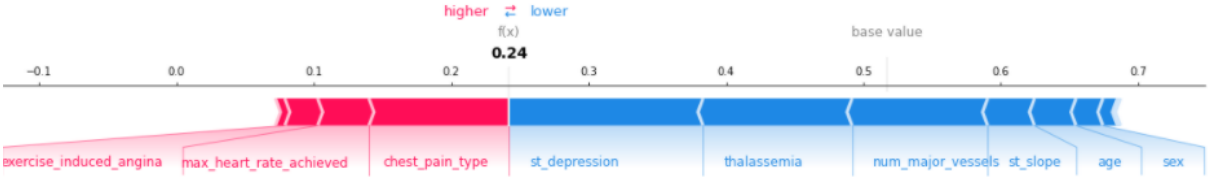

Figure 5.5: Test case :2

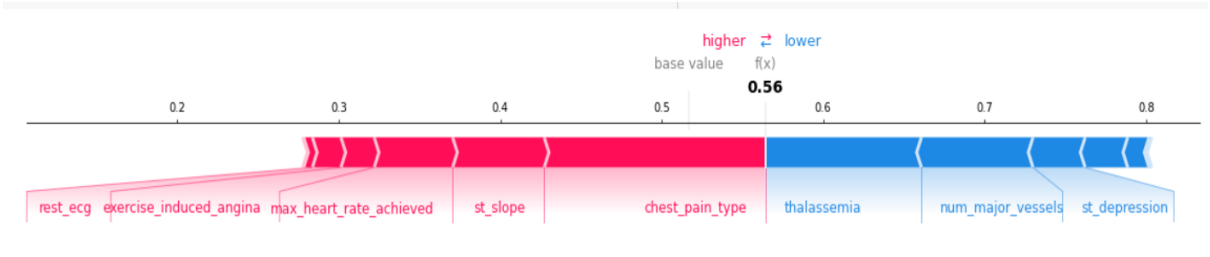

Figure 5.6: Test case :3

Chapter 6

Medical Question Answering using BERT

6.1 Introduction

This report is on the fourth objective of our major project "Integrating AI in a healthcare Web application". In this paper we create a web application that assists doctors in prescribing the best medicine.
We have used a pre trained BERT model(Bidirectional encoder representations from transformers) which is natural language processing model to find answers from set of large input documents .
Here we have used WHO documents on a list of essential medicines , essential medicines for children, cardiovascular diseases and chronic respiratory diseases.

6.2 BERT

BERT is a model developed by researchers at google which could be used for NLP tasks and Question answering tasks. Here we use a pre-trained BERT model trained on SQuAD v1.1(Stanford Question Answering Dataset)
In our application we make use of BERTs ability to do Q and A tasks. Here we provide a text sequence that is our WHO documents, and a question is asked based on these documents and the software highlights the answers in the text. The model is trained by learning two extra vectors that highlight the beginning and end of the answer.

6.3 cdQA

Closed domain Question answering is an open source software for Question answering tasks using the concept of transfer learning with a pretrained BERT model which is a PyTorch version by Hugging face(a python library that provides pretrained models). Here we use cdQA to download the bert-SQuAD_1.1 model, converts docs to pdf and pipelining.

6.4 Algorithm

- Download cdqa
- Download documents
- Convert documents to pandas dataframe
- Create pipeline for q and a
- Use pipeline to predict
- Answer and relevant paragraph output by the system

6.5 Application Structure

In our project we have used the Streamlit open source python library to demonstrate our question answering webapp. Streamlit is a frontend framework. Here we include a screenshot of our app UI (fig 1). Questions related to treatments and medicines can be typed in the box and the application would output the answer and relevant paragraph in the document.

6.6 Results

This is the screenshot of the results of the web app.
The algorithm outputs the answer as non-steroidal anti-inflammatory medicines which is the actual medicine type for gout. The app also outputs the title of the document from which the answer was taken along with the particular paragraph.

6.7 Figures

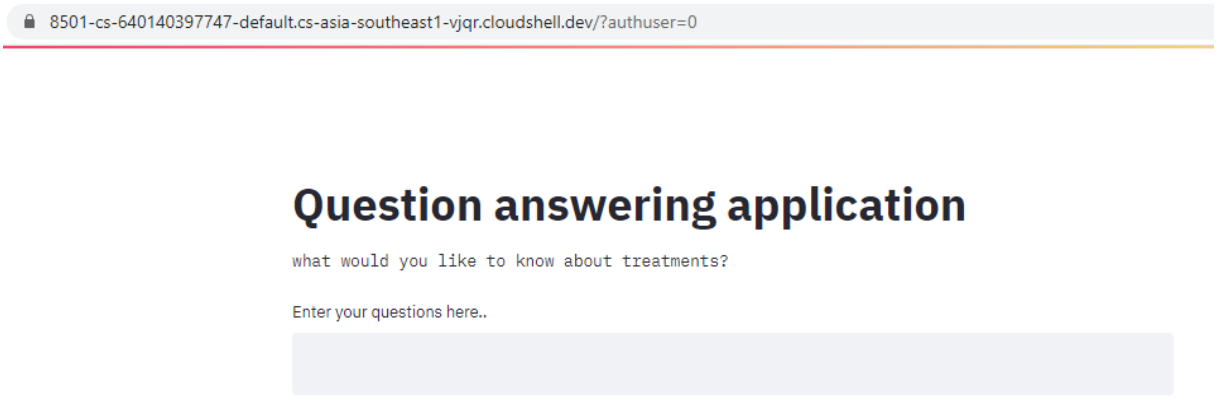

Figure 6.1: APP user interface

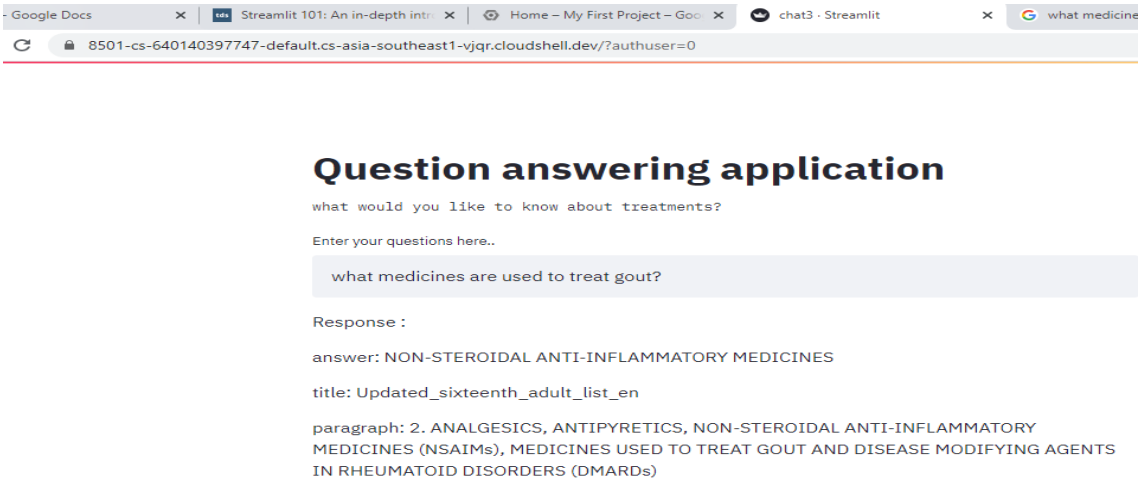

Figure 6.2: APP user interface

Chapter 7

Conclusion

Our application intends to bring the above stated objectives to effect to improve efficiency and accuracy in the healthcare sector.

The use of Magnetic resonance imaging for detection of tumors is increasing as it does not use radiation and therefore is considered a noninvasive procedure. Detection of tumors manually by a radiologist can be quite tedious and error prone. These brain tumors can cause complications, even death and therefore accurate diagnosis is essential. Our model can detect enhancing and non enhancing tumors with good values of sensitivity and specificity hence, making the model more reliable.

According to recent research, conditions such as cardiomegaly, pneumonia, etc are commonly found in patients across the globe and are the main causes for patient mortality. Therefore, an accurate diagnosis is essential in preventing further complications. These areas have a significant shortage of expert radiologists who can interpret these X-Rays effectively, hence making this process more error prone. In our model, we aim to assist such cases with more precision. Our model presents a good AUC score in detecting 12 out of the 14 major recurring chest conditions.

Around 15% of the global yearly deaths are caused due to prevalent heart diseases. Heart diseases can be sudden, making it challenging to discover as well as track the factors that contribute to the depreciation of the heart health. The ultimate aim of the prognosis functionality of our application is to provide a visual tool of a patient's current heart health. This tool makes it easier to jot down the factors that can contribute to survival risk. With an accuracy of 89.268%, the health workers can rely on the model and provide the necessary treatments.

A search engine has the ability to get results, but a lot of them become ambiguous when considering complex or conversational queries, hence BERT

is chosen as it processes words in relation to all the other words in a sentence, rather than one-by-one in order. The aim of the BERT model is to provide output for questions on a given set of localised topics. The models can also be modified for cases where data is private and only accessible to the hospital staff. And at last the model provides an accurate answer without the need to manually go through an exhaustive set of documents.

All these objectives are visualised as features on the front page of our web application. The user interface provided is user friendly, making it easy to access and switch between the features.

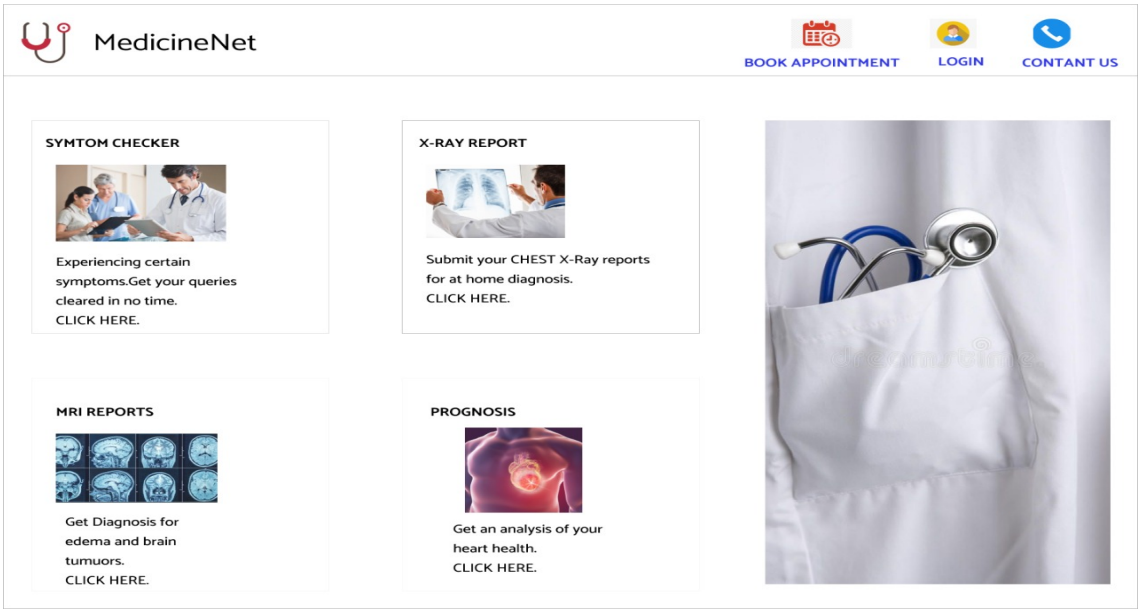

Figure 7.1: Front Page of the Application

References

[1] W. Koehrsen, "Transfer learning with convolutional neural networks in pytorch." [Online]. Available: https://towardsdatascience.com/transfer-learning-with-convolutional-neural-networks-in-pytorch-dd09190245ce

[2] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, pp. 1–17, 11 2018. [Online]. Available: https://doi.org/10.1371/journal.pmed.1002683

[3] T. S. Kumar, K. Rashmi, S. Ramadoss, L. Sandhya, and T. Sangeetha, "Brain tumor detection using svm classifier," in *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 2017, pp. 318–323.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[5] 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, 2020. [Online]. Available: https://www.mdpi.com/2075-4418/10/6/417

[6] L. Yang, H. Wu, X. Jin, P. Zheng, S. Hu, X. Xu, W. Yu, and J. Yan, "Study of cardiovascular disease prediction model based on random forest in eastern china," *Scientific Reports*, vol. 10, 03 2020.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[8] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://www.aclweb.org/anthology/D16-1264

[9] E. P. Balogh, B. T. Miller, and J. R. Ball, Eds., *Improving Diagnosis in Health Care*. National Academies Press, 12 2015.

[10] D. Greenfield and S. Wilson, "Artificial intelligence in medicine: Applications, implications, and limitations," *Harvard University*. [Online]. Available: https://sitn.hms.harvard.edu/flash/2019/artificial-intelligence-in-medicine-applications-implications-and-limitations/

[11] B. S. Vittikop and S. R. Dhotre, "Automatic segmentation of mri images for brain tumor using unet," in *2019 1st International Conference on Advances in Information Technology (ICAIT)*, 2019, pp. 507–511.

[12] S. Banerjee and S. Mitra, "Novel volumetric sub-region segmentation in brain tumors," *Frontiers in Computational Neuroscience*, vol. 14, p. 3, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fncom.2020.00003

[13] S. H. Tsang, "Review:3d unet volumetric segmentation (medical image segmentation." [Online]. Available: https://towardsdatascience.com/review-3d-u-net-volumetric-segmentation-medical-image-segmentation-8b592560fac1

[14] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *CoRR*, vol. abs/1705.02315, 2017. [Online]. Available: http://arxiv.org/abs/1705.02315

[15] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[16] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. Langlotz, K. S. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *CoRR*, vol. abs/1711.05225, 2017. [Online]. Available: http://arxiv.org/abs/1711.05225

[17] M. Almuhayar, H. H.-S. Lu, and N. Iriawan, "Classification of abnormality in chest x-ray images by transfer learning of chexnet," in *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, 2019, pp. 1–6.

[18] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[19] R. Joseph, "Grid search for model tuning." [Online]. Available: https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

[20] D. Krishnani, A. Kumari, A. Dewangan, A. Singh, and N. S. Naik, "Prediction of coronary heart disease using supervised machine learning algorithms," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 2019, pp. 367–372.

[21] S. M. Lundberg, G. G. Erion, and S. Lee, "Consistent individualized feature attribution for tree ensembles," *CoRR*, vol. abs/1802.03888, 2018. [Online]. Available: http://arxiv.org/abs/1802.03888