Multi-label classification and description generation of Pulmonary diseases in Chest X-rays using Deep Learning techniques

MSc Research Project
Data Analytics

Madhav Kant Lavania
Student ID: x18104151

School of Computing
National College of Ireland

Supervisor: Mr. Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Madhav Kant Lavania
Student ID:	x18104151
Programme:	MSc. Data Analytics
Year:	2018-2019
Module:	MSc Research Project
Supervisor:	Mr. Vladimir Milosavljevic
Submission Due Date:	12-08-2019
Project Title:	Multi-label classification and description generation of Pulmonary diseases in Chest X-rays using Deep Learning techniques
Word Count:	8926
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to the research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Madhav Kant Lavania
Date:	12-08-2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multi-label classification and description generation of Pulmonary diseases in Chest X-rays using Deep Learning techniques

Madhav Kant Lavania
x18104151

August 12, 2019

Abstract

The applications of deep learning have broadened their spectrum in the field of medical research. One such field is medical radiography which uses several imaging techniques like CT-Scan, X-Ray to view the internal abnormalities of a human body. Traditional image classifiers could not process these noisy, blurred or unclear X-Ray images, which leads to incorrect results. These image classifiers lack in mimicking the exact understanding of a professional, which comes from rigorous training and hands-on experience. Hence, a novel solution to x-ray image classification and description generation has been presented. This approach works towards detection and classification of pulmonary diseases like fibrosis as well as generates a medical interpretation of the detected abnormalities. This model is prepared in two parts. The first part works as a multi-label classifier and is based on a pre-trained MobileNet Convolutional Neural Network (CNN), while the second part of the model is designed by combining a pre-trained CNN VGG16 and a Long-short term memory-recurrent neural network (LSTM-RNN) to generate a medical description of the diagnosis. Chest x-ray image dataset from OpenI and NIHCC with their medical labels have been used for training and testing the model. This multi-label classifier and the captioning system performs well with an overall prediction accuracy of about 87% and BLEU score of 0.58.

Keywords: Medical Imaging, Deep Learning, CNN, MobileNet, VGG16, LSTM-RNN, Word Embedding, Image captioning

1 Introduction

Interstitial lung disease (ILD) is prevailing rapidly in many parts of the world with a huge increase in the number of people suffering from it[1]. ILD is a group of about 250 different lung diseases characterized by damaged and scarred lungs. Prolonged illness due to ILD may result in pulmonary fibrosis, which is again a chronic lung disease identified by scarring of lung tissues. These scars become prominent over time and ultimately hinder the respiratory process of the infected individual (Anthimopoulos et al. (2016)). Early screening of such diseases can do damage control.

Physicians generally ask for getting a chest X-ray done before recommending any other major high-resolution imaging like CT-scan or MRI. X-rays are cheaper and more easily available than their high-quality counterparts. Doctors can easily diagnose the problem by taking a quick scan of the X-ray. However, due to the shortage of experienced radiologists and an ever-increasing number of patients, it is becoming challenging to diagnose Pulmonary fibrosis and other such ILD diseases. X-rays are the preferred source of identifying the problem because of their low cost and easy availability, even in rural areas. Since X-rays bear a lower quality resolution, sometimes due to which few diseases might not get correctly identified by the radiologists which further results into critical health issues or even death of the patient. Thus, a classification

[1] https://www.who.int/gard/publications/The_Global_Impact_of_Respiratory_Disease.pdf

system is required which can not only identify these diseases efficiently but also generates related diagnosis description of the same.

Therefore, in this research, we put forth a multi-label based classification and report generating a model which works specifically for chest X-ray diseases including Pulmonary Fibrosis using the X-ray images and their diagnosis reports. Many conventional machine learning models like support vector machine (SVM) have been implemented in the past for identification of these diseases by highlighting the affected area with the help of bounding box outlining the problems. Computationally, creating bounding boxes is a time-consuming task and these algorithms are also slower in performance which may behave very differently in case of larger datasets. The applications of deep learning approaches have been applied in areas like text analysis, image classification and captioning, Natural Language processing. These algorithms are meant to be used with larger input data. Multiple past researches like (Krizhevsky et al. (2012), Chen et al. (2017*a*) show that the combination of CNN and LSTM-RNN may be perfectly used together for language translation, image captioning, and Natural Language Processing. This research presents a solution which not only identifies/detects the lung-related diseases like Pulmonary Fibrosis but also generates a medical description of the same. The model has 2 major parts. The first part is a multi-label classifier which uses a pre-trained CNN "MobileNet". The weights from this part are inputted to the second part of the model which is a combination of another pre-trained VGG16 CNN and an LSTM-RNN network. This second unit works as a disease description generator. The combination of both of these together works best as a multi-label disease detection and captioning system.

MobileNet CNN is a lightweight, highly accurate and a faster CNN designed by Google. It has been pre-trained with thousands of embedded images. This type of CNN is perfect for a noise-prone, dirty data with blurred input images.

The second part of the model combines the results of classification from the first part. Its weights are given as input to another pre-trained CNN model VG16 which has been trained on about 15 million image-label based combinations. The idea here is to leverage the use of these pre-trained classifiers to achieve the highest performance in terms of ILD detection and description generation. Implementation of LSTM facilitates the report generation as it works best for sequential data. The input to LSTM is the normalised sequence of numerical data in the form of a vector. So, a word embedder is used to convert the image labels into a feature vector. This vector set is then provided to the LSTM which generates the respective description word-by-word for a given image after its correct classification. A special inclusion of blurring, flipping, shearing, rotating, augmenting and tilting of the images has been done to enhance the performance of the model. This research is completely based upon the chest X-ray images and their respective labels. Two different data sets have been used in this study. First one is from NIHCC (Wang, Peng, Lu, Lu, Bagheri & Summers (2017)) which a public data repository, containing about 112,000 anonymised chest X-ray data with information of about 14 different pulmonary and ILD diseases. The second dataset is taken from OpenI (Demner-Fushman et al. (2015)) which is again a publicly available anonymised dataset which is accessible at URL [2]. It contains approximately 7470 chest X-ray images and 3,955 different radiology reports.

In the upcoming sections, vast literature review is presented in section 2. Methodology used in this research is described in section 3.4, followed by Implementation and Evaluation in section 4 & 5. The paper is concluded with some discussion on results, conclusion and future work in section 6 & 7.

2 Literature Review

For decades, machine learning algorithms like K nearest-neighbour, random forest and support vector machine were handed-down for general classification problems. But, these algorithms

[2] https://openi.nlm.nih.gov/gridquery?q=chest%20x-ray%20fibrosis&it=xg&sub=x&m=1&n=100

need a bounding box across the area that needs to be identified. Creating bounding boxes is a hectic and time-consuming task which requires human intervention and good specific domain knowledge of medical imaging. Additionally, tradition algorithms are slow and usually take days for their training. Moderate performance, less accuracy, and weak classification power are few of the factors that push to research about new technologies.

These problems are resolved by deep learning which turns out to be an exceptional solution in labelling, annotating, classifying and detecting a single or a multi-label problem. Below are few researches performed on image classification and captioning using deep learning techniques such as CNN and RNN.

2.1 Deep Convolutional Neural Networks in Pulmonary disease classification

Research by Anthimopoulos et al. (2016) used chest X-rays followed by a CT scan for classifying pulmonary interstitial diseases. This paper suggests that to infer some insights from an image, the output of a fully connected layer should be stored in a single vector. This helps in proper image classification. An image feature should always be equal or small to the size of the texture, otherwise the chances of transferring inappropriate information increases. Hence, a model with 2*2 kernel size having 5 convolutional layers was designed.

'LeakyRelu' activation function was applied to every convolutional layer along with an average pooling layer. A dropout layer was implemented with Relu as an activation function to prevent the model overfitting. For training & generating the weights, Adam Optimizer and cross-entropy loss function were added. The learning rate was set to 0.001. The dataset of 120 images was converted into 15K small patches, 32*32 in size, for rigorous training. Later, with the help of these functions, the model designed was run several times by changing few functional parameters. However, the model showed slow training because of its small learning rate and a large number of parameters. Overall, the concept of LeakyRelu helped in transferring important and non-zero information to the fully connected layers, which set an inspiration for our work.

Salehinejad et al. (2018) dealt with a small and imbalanced dataset for classification or detection problems by artificially generating them. The researcher introduced a generative adversarial network (GAN) which generates the data identical to the raw dataset. A CNN model is then trained on both real and artificial images to have a big dataset for the better results. A combination of Generator and Discriminator is applied. The generator takes an image as an input, process them by mapping all the features, and generates a stimulating image out of it. This stimulated image is then passed through the discriminator to compare and check the matching of it with the raw images. Lastly, all the images (generated and raw) are used for training the CNN. GAN helps in improving the quality, efficiency, and accuracy, on the other hand, eliminates the possibility of unbalancing and over-fitting. The deep CNN (DCNN) that author has used comprises of 5 convolutional layers with 3 fully connected and max-pooling layers. Whereas, each part of GAN had 6 convolutional layers each.

A brilliant model designed by Rajpurkar et al. (2017) contained 121 convolutional layers called dense convolutional neural network (DenseNet) to detect Pneumonia in lungs using chest X-rays with the best accuracy and results. The analysis was performed on a recent publicly available chest x-ray dataset called ChestX-ray14. Pneumonia has a very vague vision because of veins and tissues. Thus to get a clear reflection and to reduce to possibility of incorrect classification a DenseNet was proposed. In the pre-processing, the chest x-ray images were converted to heat maps which is useful in highlighting the affected area. Comparatively, it then becomes easier for CNN to understand the pattern of the highlighted area. The objective of using such dense layers was to calculate the optimisation of the model at every step. A separate dataset of about 400 images was collected, this dataset was labeled by 4 different doctors from Standford university. It was later noticed that the F1 score attained by the labels of experienced doctors was comparatively low than the score achieved by DenseNet. This reflects the power and capability of deep learning techniques. Considering the images in multi-scale would have bought more weight and importance to the experiment. As the dataset was huge, still the model was only capable of classifying Pneumonia at a time. Adam Optimizer and cross-entropy loss function were used for training the model.

To improve performance and to classify 14 different pulmonary diseases, Xu et al. (2018) extended the experiment conducted by Xu et al. (2018) using the same NIH dataset that contains 112K chest x-ray images. The dataset was completely analysed here and the problem of unbalancing has been highly taken into consideration. A pre-trained and one of the most prominent CNN model InceptionV3 is used to classify pulmonary diseases. The last layer of the model is replaced by 15 perceptrons. Each perceptron is defined as a specific class. 14 are for different disease and 15th signifies 'no findings'. Hence, each perceptron gives the chances of specified disease in percentage and accordingly a perceptron with the highest percentage is the classified disease.performance and to

Summary 1: Researchers in the above experiments have only focused on classification or detection of a single disease at a time, instead of focusing on multi-labeled diseases. A patient can have more than 1 disease at a time, finding one and neglecting another could be very dangerous. It is quite difficult for CNN to understand an image if it gets rotated, thus, to work with medical imaging it is necessary to consider all the limitations and work upon it. However, one of the researchers has trained their models with rotated or in multi-scale images. Unavailability of large datasets is also a major concern. Other than these, all the models were capable of achieving a good accuracy which shows the success of deep learning in the field of medical imaging.

2.2 Multi-scale, rotated, heat map, and multi-resolution medical images

X-rays, MRI's and CT scans are sometimes not perfect and may require some noise clearance before sending it to the doctor for diagnosing the problem. The machines get old and pathology labs would require a huge amount of money to buy a new one every time the CAD images get distorted. The researches discussed above have performed well in classifying a single-labeled image at a time, but they have taken images on a different scale, rotated images and even the images in different resolutions. Thus below papers have considered images in different scales for classification and segmentation problems.

Shen et al. (2016) explored the use of multi-scale images to detect lung cancer with the help of nodules using CT scan radio images. The author used two different datasets and developed a model in two different parts. The first part uses 2272 lung nodule (swelling) images that are not diagnosed with lung cancer. The CNN performs feature extraction and then transfer it to MIL. Multiple Instance Learning (MIL) is the second part of the complete model. It comprises of 3 different CNNs. These CNNs have been trained in different scales using 150 lung cancer diagnosed images. Extracted features are then matched with the patterns available in the second part to conclude binary results. Even if one of the featured matches, the image can be concluded as positive with lung cancer otherwise not. The combination of CNN-MIL performed well but, 'regression' is used as a loss function which is not ideal for image classification.

Another research for the classification of skin lesion (tumor or burned areas) was performed by Kawahara & Hamarneh (2016) using an already trained CNN 'AlexNet'. The study is based on considering the images in multiple resolutions to rectify the chances of misclassification by augmenting the images into multiple resolutions. Sometimes the affected area is very small and can't be traceable with naked eyes. Thus, to cover every area it is important to consider the images in different resolutions. A similar experiment has been carried out by Sermanet et al. (2013), where a model was trained with multiple resolutions but the mapping of features was inaccurate and in another test using a different resolution, model's performance eventually got degraded. There are other experiments where the models were tested on different resolutions and later aggregated the output but were not trained on them. These issues lowered the classification performance and failed to establish a connection between training and testing. Thus, Kawahara & Hamarneh (2016) worked on these problems by developing two tracts. These tracts produce the image in the form of m*n*4096. High-resolution images were tackled by the lower tract and low to the upper tract. Training time was reduced by eliminating the fully connected hidden layers of AlexNet. The output in the form of a single vector was used in skin lesion detection and classification.

Lung nodule detection with the help of multi-view CNN was presented by Setio et al. (2016). The model uses CT scan images as the input dataset. The design not only works in detecting nodules but also reduces the false-positive occurrences. The model was designed in three different parts which work in a multi-stream fashion. First part follow the approach developed by Murphy et al. (2009), Jacobs et al. (2014) and Setio et al. (2015) to detect the nodules and to covert the image into multiple patches. By clustering the nodules in 3 different classes (solid, subsolid, and large solid), these researchers were able to ease the classification problem. The output of these classes is then converted into a single feature vector that defines them properly. These authors have defined a threshold which helped in reducing the false positive rate. However, there can be scenarios where the negative class will be classified as positive. In the second part, the patches are converted into 9 different scales and each type of scale is dedicated to a specific CNN. It was observed that the accuracy got improved by 6% by eliminating false positives and by leveraging multi-scale CNN.

Wang, Zheng, Yang, Jin, Chen & Yin (2018) presented a multi-stream CNN to classify five different lung textures. As CNN can't understand or able to classify a rotated or tilted input if it isn't initially trained on it. The author understood this issue of rotation-variant and tilted all the CT scan images to a random scale. Then, with the help of a Gabor LBP (local binary patterns) the rotated images are converted to a systematic format. As per the paper, this technique has proved to be the best prominent way to deal with the problems of rotation variant. Images are augmented and cloned which belongs to the minority classes to overcome the issue of imbalance. Now, a CNN model was prepared to extract a feature vector using the balanced dataset. The model's overall performance got increased by using Gabor LBP and by eliminating the problem of rotation-variant.

Extending the above research, Wang, Zhou, Gevaert, Tang, Dong, Liu & Tian (2017) has taken an axial, coronal and sagittal view of CT scan images to segment lung nodules using a multi-stream multi-scale CNN. The dataset with high-resolution images was converted to small patches to identify and classify an image. To stay away from overfitting, 1-norm regularisation that helps in removing zero or very tiny values is used. The use of Xavier algorithm is defined as the heuristic in the complete model. This algorithm is used to load the weights in the form of a feature vector, which is then updated through stochastic gradient descent algorithm. The only limitation that the author has highlighted is the use of small dataset and desire to use the same approach on a bigger dataset as well. Results showed improved performance compared to similar conventional researches.

Summary 2: In this section, researchers focused upon classifying a disease or a pattern at a given time. These researches could have been complete if all three; multi-scale, multi-resolution and rotated images were considered at once. Researchers have also majorly focused on classifying one disease or pattern at a time and the problem of multi-labeling remained intact. It can also be seen that major experiments were done on CT scans. None of them have worked on Chest x-ray which is the most promising initial step of diagnosis.

In the field of human-action recognition (Tu et al. (2018)), video-classification (Wu et al. (2016) & Wu et al. (2015)), gesture recognition (Wei et al. (2017)) and in-text writer identification (Xing & Qiao (2016)), the use of augmented images is becoming very popular. All these experiments have used deep learning model with different functions depending on their data.

2.3 Combining CNN and RNN to address multi-label issues

Multi-labeling or image captioning is another major issue. Below the papers have considered multiple labels and caption generation to describe images:

2.3.1 Classify multi-label pictures or videos and generate captions in non-medical fields

A wonderful study performed by Bai & An (2018), analyses the issues of captioning and annotating the images or videos and also compared different methodologies that have been used for captioning tasks. Bai & An (2018) highlights the fact that a single image can have multiple labels hidden into them. Therefore, it is important to extract these labels and format them as sentences. A sentence representing the image becomes a caption.

Medical imaging is a field where doctors have to be careful before giving their decision and making a report. This becomes tedious when there are many people and only 1 radiologist or doctor among them. Sometimes, this could also lead to misdiagnoses of the primary disease and creates more problem for patients. Thus, researchers are trying to figure out a way which could generate a report automatically and give important insights using a machine. While comparing various methodologies, thus, deep learning encoder-decoder techniques are the best suited for a similar task.

Model by Vinyals et al. (2015) takes images as input and generates a caption for it. The combination of CNN-LSTM not only helps in classifying the image to its correct class but also annotates it. Fully connected layers of CNN has been removed and the output of the previous layer is directly fed to a pre-trained LSTM. LSTM contains 3 input gates, so, the output of CNN is fed through one and textual data through another. The experiments use Stochastic gradient descent which helps in reducing the loss and updating the weights. The whole model was evaluated using the BLEU score. It compares the predicted caption with the actual and then generates the score. Size of the dataset plays an important role in improving the BLEU score. In initial testing the score came out to be 27% and when the data set got increased the score improved to 49%. Donahue et al. (2015) has extended the above research by implementing the same technique to generate captions for videos.

Chen et al. (2017*b*) has also worked on the encoder-decoder technique using textual data only. The objective of this research is to capture all the annotation that can be inferred from the image and categories them. A tokenizer 'word2vec' has been used to convert the text into the desired format to fed it to CNN which converts the output of word2vec into a feature vector. The output generated post passing the features through fully connected layers is fed to RNN as an input with the help of linear transformation. This ensemble technique helped in removing the complexity and improving the computational power and overall efficiency of the model.

2.3.2 Classify multi-label images and generate captions in medical imaging

In the medical domain, researchers have started leveraging deep learning encoder-decoder technique for dealing segmentation, detection, annotation, and contextual labeling problems. Below are a few of the studies based on a similar approach.

Attia et al. (2017) developed a CNN-RNN model for segmenting the flow of surgical instruments. These days, no surgery is being done without machine involvement. Having a clear vision of the operating area is quite impossible with bare eyes, there are veins, blood, and tissues that cover the affected area. Thus, before any operation, the surgical tools are trained and it works accordingly. But there have been scenarios wherein due to some ad-hoc problems (shadow, blurriness) it becomes impossible for the tools as well to move further with it. Thus, to eliminate such problems in critical situations, the cameras fitted on the tools send an image or video to CNN. CNN then clear the noise from the image and accordingly send the information to RNN which helps in contextualizing the labels and instantly train the tool and guide the surgeon to move ahead with it. In such a critical task, it is also our duty to not trust machines completely, thus, according to how the tool is trained sends images continuously, CNN converts it into a feature vector and send it to RNN, RNN then matches with whatever is displayed in the image and on what it is trained and throws and alert to the doctor. It is a complex task and the researcher was able to acquire a decent result while evaluating the performance of the overall model. Whereas Cai et al. (2017) worked on segmenting pancreas to infer true insights from CT scan images.

Wang, Peng, Lu, Lu & Summers (2018) have used attention-guided approach for classifying and describing chest x-rays. The difference in attention-guided (A-G) and encoder-decoder are that the A-G uses pre-trained RNN. A saliency Weighted Global Average Pooling (SW-GAP) layeris replaced with the last layer of CNN whose output is fed to RNN. Both LSTM-RNN and CNN were pre-trained. Results were calculated using BLEU score and the score that this research has acquired is lower than the previous researchers. Also, this research doesn't consider different scaled, rotated or different resolution images.

In another paper Shin et al. (2016) have used an already trained deep learning model for labeling pneumonia and thorax problems. The heuristic in this research is the use of batch normalisation and data drop regularisation techniques for training the model. Two types of RNN; LSTM and gated recurrent unit (GRU) have been compared. It was seen that using the batch normalisation and LSTM deep learning method performed better than GRU using data drop regularisation.

Summary 3: The common problem of multi-class labeling highlighted in summary 1 and 2 has been using an encoder-decoder technique, however, the use of multi-scale, multiple resolutions and rotated and heat map images has not been considered in the researches discussed in this section. Also, none of the research has taken pulmonary fibrosis as their primary classifier which is indeed a deadly disease and can give birth to other pulmonary diseases as well. Overfitting which is another major concern while using the encoder-decoder technique. It can be concluded that by avoiding a few parameters and using a bigger dataset, such problems can be eliminated. Thus, we can say that the use of the augmented image for the generation of the description using the encoder-decoder technique has not been explored as of now.
Therefore, this gives birth to our research question "How effectively and accurately can pulmonary fibrosis along with other pulmonary diseases be classified and described using an encoder-decoder technique?"

Improvement: Inspired by all the researches discussed, a novel holistic model has been designed in an encoder-decoder technique. The model outperformed in terms of predicting multiple diseases from an x-ray (if exists) and is also able to generate a description of an image. Details of the methodology, modelling, evaluation, and results are illustrated in the following sections.

3 Methodology

Two major methodological approaches are majorly followed to streamline the modelling process; CRoss Industry Standard Process for Data Mining (CRISP-DM) and Knowledge Discovery Databases (KDD). (Azevedo & Santos (2008)) performed a comparative analysis between the two and concluded that the stages of CRISP-DM is not repetitive and provides a full understanding of the research and its objectives. Thus, CRISP-DM has been used as our research methodology. The architecture and all the phases are shown in figure 1.

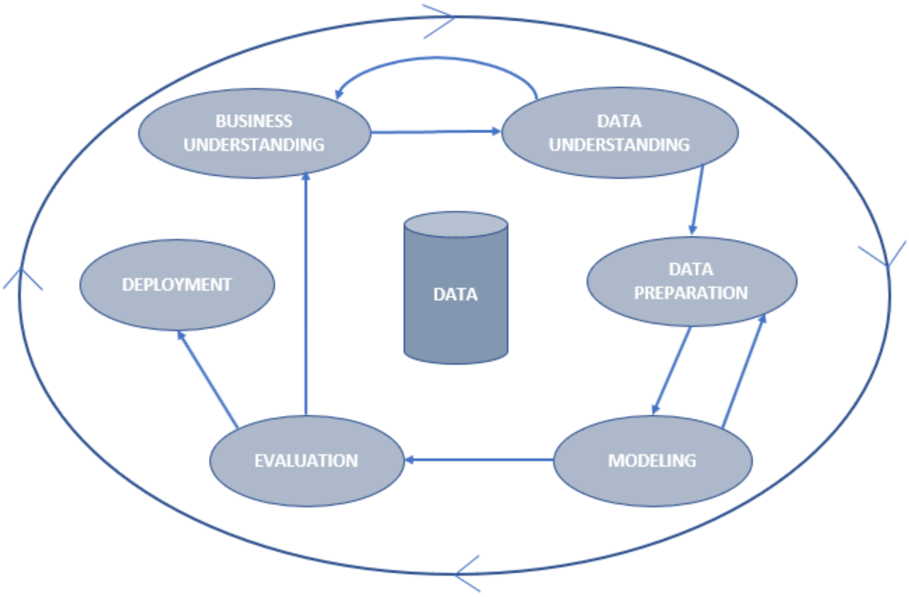


Figure 1: CRISP-DM methodology

3.1 Business Understanding

Before working on any project it is important to gather its complete objective. The aim is to produce a model which detects pulmonary diseases including pulmonary fibrosis and prepare a report that covers the predicted percentage of a problem that is affected and its impression using x-rays which are in rotated, have multi-resolution or are in multi-scale format. This problem has not been addressed in this way before. The researchers who have tried captioning the CAD images haven't considered fibrosis in their work. Their work has gained proficiency in the classification part but still struggle to achieve a better result in generating auto reports. The target people are ones belonging to rural areas who cannot afford expensive CT scans or MRI. This model will be a major advantage to the people who can't afford expensive treatments and multiple visits to the doctors. If the report comes out to be negative, the patient doesn't need to visit the doctor again.

3.2 Data Understanding

No model can be prepared without a complete understanding of data. It's quite impossible to perform data modelling without a proper understanding of its components. Data understanding also means understanding the sources of it. Data can be available at multiple websites but it isn't ethical to use them. Or if all the necessary permits are present to access it, then the source sometimes isn't reliable. Thus, it is important to acquire data from a reliable source.

For this research, labeled chest x-ray images and reports written by doctors were needed. Datasets which have been finally used are as follows:

- **National Institute of Health (NIH):** NIH has recently released a dataset containing 112K chest x-rays with their labels. It is one of the largest available dataset belonging to medical domain. It contains real-time chest x-rays of about 31000 anonymised patients and all the personal information have been anonymised. The dataset is designed in such a way that it can become useful not only for deep learning but also for traditional image classification algorithms like SVM. Separate document to build data understanding along with an excel file containing labels of all these images has also been provided. Full dataset and its description can be downloaded from URL[3] & [4] [5], Wang, Peng, Lu, Lu, Bagheri & Summers (2017).

- **OpenI:** It has been the most common medical image repository for the researchers. OpenI is a different platform but owned by NIH only. Researchers have just used its images for image classification however, none of them have used their report features. Along with medical images, OpenI contains pathology reports which provide deep insights about the disease, severity of it, etc. The dataset is completely anonymised and the source provides full access to it. This dataset contains 7470 chest x-ray images and 3955 unique radiology reports. Datasets can be downloaded from OpenI website[6].

3.2.1 Ethics

Both the datasets are publicly available and can be used for non-commercial purposes. All the images and reports are anonymised and it is impossible to gather a patient's personal information from it.

3.2.2 Data Exploration

The best way to explore or understand the data is with the help of visualisations. Visual diagrams are easy to understand and gains user attention. Therefore, the findings of NIH dataset is as below:

- Disease distribution with respect to gender is shown in figure 2.

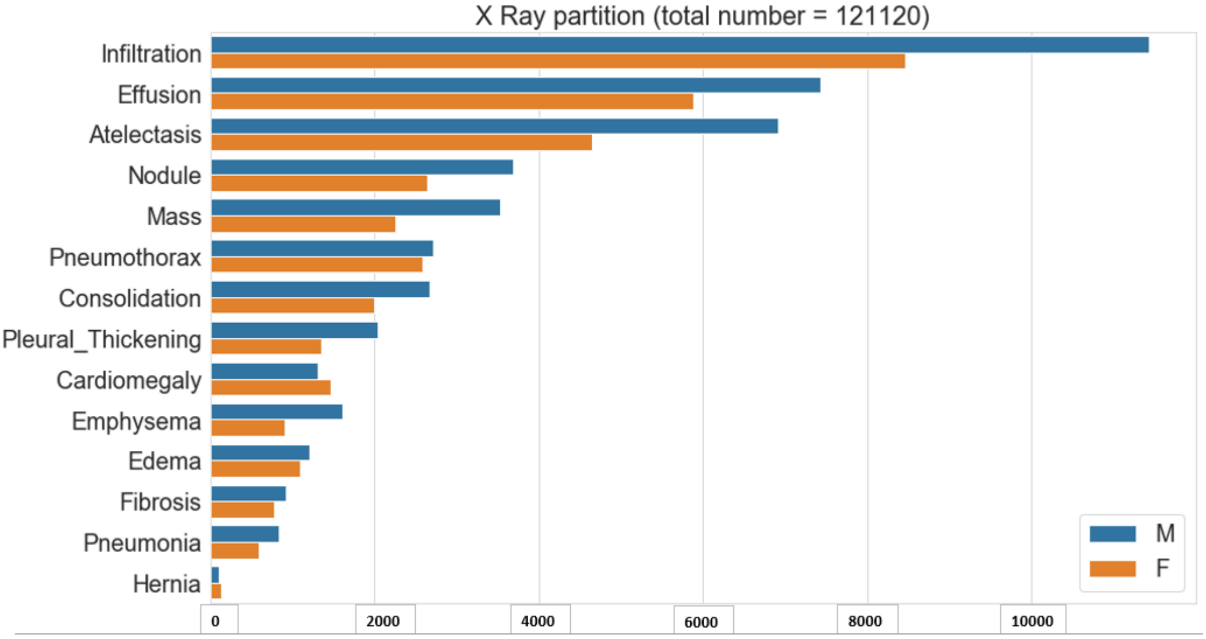


Figure 2: Chest X-ray partition with respect to diagnosed diseases and gender

- Comparison between the x-rays containing simple and multiple diseases with one main problem is shown in figure 3.

[3] `https://nihcc.app.box.com/v/ChestXray-NIHCC`
[4] `https://cloud.google.com/healthcare/docs/resources/public-datasets/nih-chest`
[5] `https://www.kaggle.com/nih-chest-xrays/data`
[6] `https://openi.nlm.nih.gov/faq#collection`

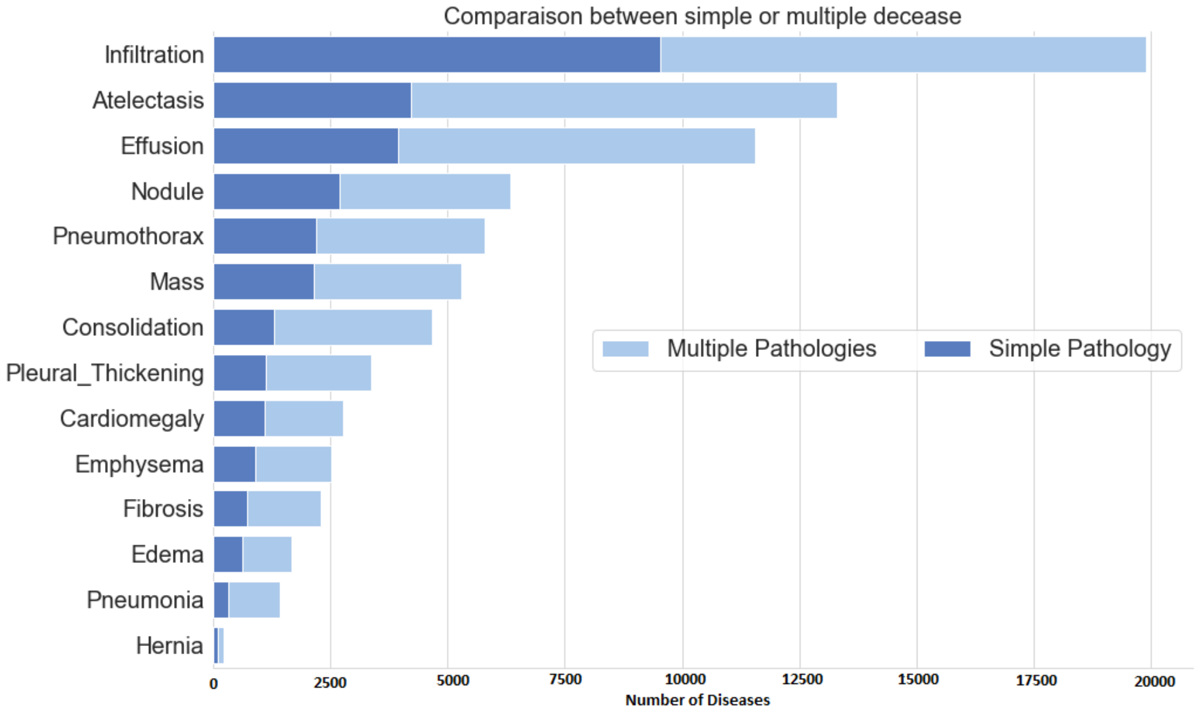


Figure 3: Comparison between the x-rays containing simple and multiple diseases

- Multiple diseases exits but the main disease is Fibrosis as per illustration 4.

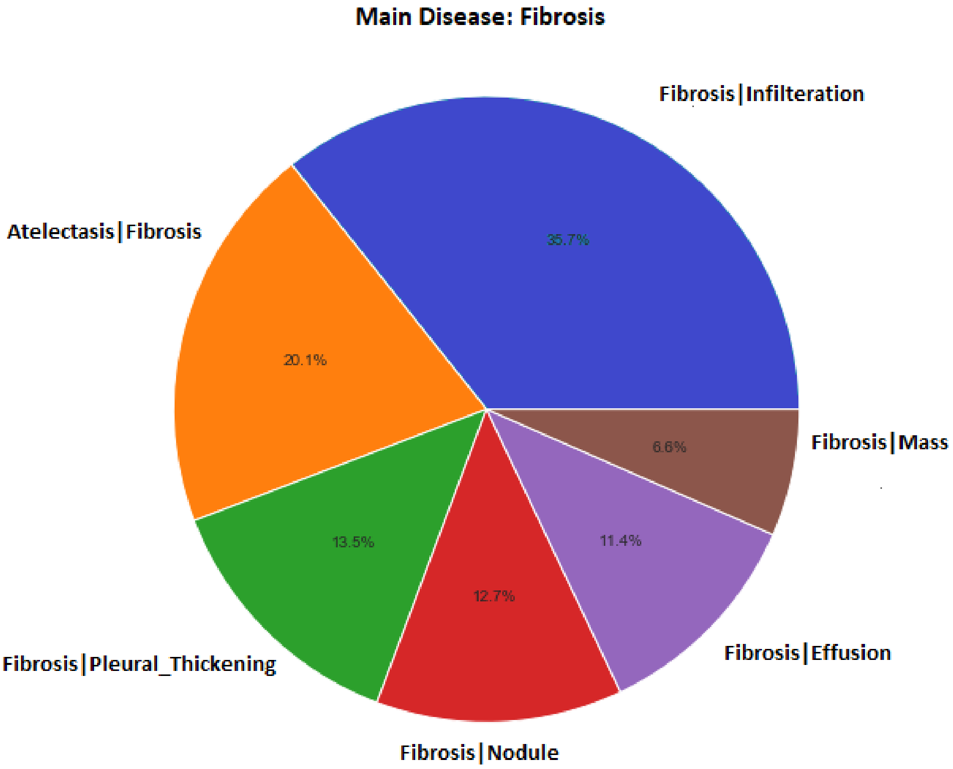


Figure 4: Fibrosis with other diseases

3.3 Data pre-processing & preparation:

Data preparation is a key step in the direction of quality decision making, improving effectiveness and offering a model a competitive edge[7]. There are always noises in big data. Possibility of human errors in terms of special characters, blank spaces, NAs, numeric keywords, URLs, and words in different languages makes it difficult for any model to understand and also reduces the performance of it. Therefore, before processing the data it becomes a completion to deal with such issues. On the other hand, there are pre-requisites for every algorithm that needs to be considered to attain good results. Thus, data has to be prepared which can be understood by the algorithm. All the pre-process and data preparation tasks are visually represented in Figure 5 along with its description as follows:

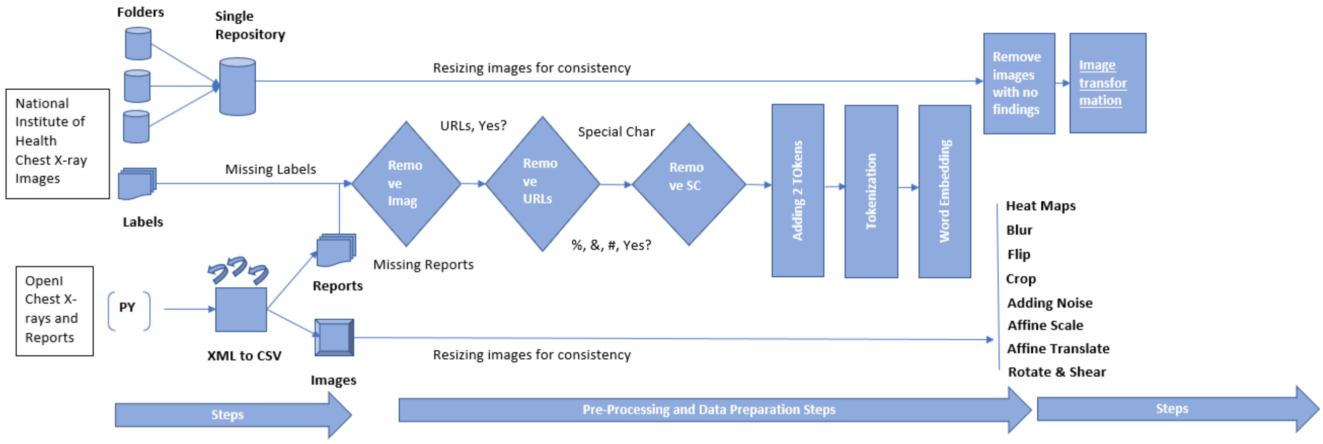
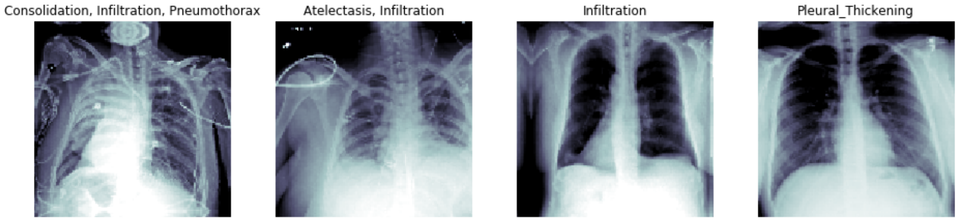


Figure 5: Data preparation & pre-processing

- **Dealing with NA's or missing data:** There are a few common pre-processing steps that are mandatory to take in every project. The data present in the CSV is checked for missing labels or NAs, labeling medical images require a lot of experience and intervention of doctors which is not possible. Thus, the images with missing labels were removed.
- **Extraction of desired data from XML to CSV:** Reports downloaded from OpenI were in XML format and to use it for captioning it has to be converted in an excel format and later passed it to the model. Python code was written to extract only the details that are needed for future use.
- **Removing xxxx:** In x-ray reports, there was the personal information of patients before making it anonymised. In this task, the words were replaced by 'xxxx', which makes no sense. Thus, before passing the data to LSTM, all the 'xxxx' were removed.
- **Converting data into simple binary format:** Performing one-hot encoding is a necessary task when dealing with a multi-class problem. This helps a machine learning model to do a better job.
- **Transformation of x-rays provided by NIH:** Due to a large number of x-ray images, it was becoming a very tedious and time-consuming task to grayscale, re-scale, zoom, rotate, shift, and flip one by one. This task would have increased the data size by 6 times. Therefore, to avoid it, all the images were passed to a function which augments all the images and then passes it to the CNN without making the task more complex. The output of the images are shown in figure 6 This was possible only because the size of the data is already huge.

[7] https://www.talend.com/resources/what-is-data-cleansing/



Consolidation, Infiltration, Pneumothorax
Atelectasis, Infiltration
Infiltration
Pleural_Thickening

Figure 6: Transformed chest x-rays from NIH

- **Adding tokens in reports:** To make the model understand about starting and ending of the sentences, startseq and endseq are added in the beginning at the end. Below is one example:

  Original text: Both lungs are clear and expanded.
  Text: startseq Both lungs are clear and expanded endseq.

- **Tokenization:** Captioning works word by word. Every word is assigned a whole number which is termed as a token.

- **Vocabulary Creation:** Words that are appearing frequently are stored in an array and a unique token is assigned to them. The frequency has been taken as 3. The reason is to get rid of outliers.

- **Word Embedding:** All machine or deep learning techniques understand numeric data. Data in the textual format is converted to numeric before feeding it further. CNN takes an image as an input, convert it into a feature vector and feeds its fully connected layers with it. Identically, a token is assigned to every unique word and a feature embedding matrix is later created using these tokens. This complete process is illustrated below.

  As discussed above, all deep learning models takes input in the form of a feature vector. Dense convolutional layers of CNN converts an image into a vector and pass it to its pooling and fully connected layer. Similarly, there is a necessity to convert textual data to process it through LSTM-RNN. LSTM doesn't have a structure like CNN which builds a vector within themselves, therefore, a vector of textual data needs to be prepared outside and model and then transferred to it, as below:

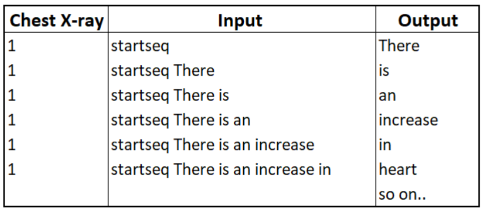
  Post tokenization, an example from reports is taken for explanation purposes:
  Report 1: startseq There is an increase in heart size endseq.
  Post tokenization, words are allocated a token: {startseq- 1, there- 2, is- 3, an- 4, increase- 5, in- 6, ...and so on }.

  Recurrent neural networks work on sequential data. For prediction, it takes one word process it, generates a possibility of another word and similarly the third word and so on. From one gate it takes a feature vector of an image and from another vector of a text.
  Input = Chest X-ray (vector) + 'startseq'; output = There;
  Again, Input = Chest X-ray (vector) + 'startseq There'; output = is ;

  A matrix is prepared based on the process and is shown in figure 7:

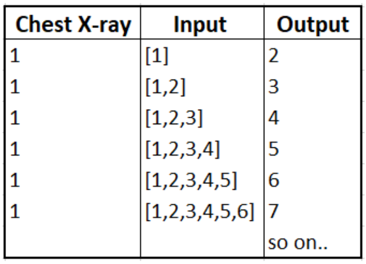
  A data point is a single line from the figure 7. An image can have more than one data points depending on the length of its caption. Now, the tokens assigned to the words while creating a vocabulary takes the place of words and the output is shown in figure 8.

  Now, this process is done using 1 sample caption of an x-ray. This one caption made 8 data points and now imagine how many data points will be created for about 8000 images?. This requires a good computation and processing power, that can only be provided by the



Chest X-ray	Input	Output
1	startseq	There
1	startseq There	is
1	startseq There is	an
1	startseq There is an	increase
1	startseq There is an increase	in
1	startseq There is an increase in	heart
		so on..

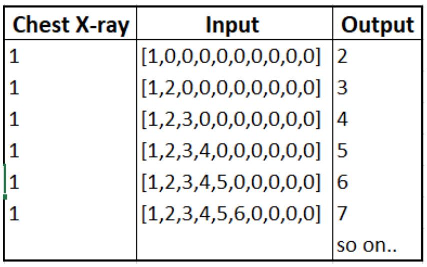
Figure 7: Feature matrix pre-processing Part 1



Chest X-ray	Input	Output
1	[1]	2
1	[1,2]	3
1	[1,2,3]	4
1	[1,2,3,4]	5
1	[1,2,3,4,5]	6
1	[1,2,3,4,5,6]	7
		so on..

Figure 8: Feature matrix pre-processing Part 2

GPUs. Thus, the training of the models was done in batches. One requirement of training the model in batches is that it needs the data points of equal length, therefore, the table as shown in figure 8 is converted into the matrix shown in figure 9, the maximum words in a caption is assumed as 10.



Chest X-ray	Input	Output
1	[1,0,0,0,0,0,0,0,0,0]	2
1	[1,2,0,0,0,0,0,0,0,0]	3
1	[1,2,3,0,0,0,0,0,0,0]	4
1	[1,2,3,4,0,0,0,0,0,0]	5
1	[1,2,3,4,5,0,0,0,0,0]	6
1	[1,2,3,4,5,6,0,0,0,0]	7
		so on..

Figure 9: Feature matrix pre-processing Part 3

This whole process has been followed by various researchers ((Vinyals et al. (2015), Karpathy & Fei-Fei (2015), Tanti et al. (2017)) and experimenters ([8], [9]) for imaging captioning tasks in different areas. The matrix shown in figure 9 is converted in a feature vector using the embedding technique before feeding this to LSTM-RNN.

- **Eliminating 'No Finding' X-rays:** NIH dataset contains about 70K x-rays that belong to the class of 'No finding'. Considering a large number of samples from this class would have overfitted the model. Thus, only 10000 of such chest x-rays have been taken for further processing.

- **Converting chest x-rays into heat map:** Inspired by Rajpurkar et al. (2017), chest x-ray images from OpenI datasets are converted to heat maps which have helped in reflecting the affected area.

- **Augmenting chest x-ray image from OpenI:** This dataset is not big as compared to the one published by NIH. OpenI have about 7000 chest x-ray images, thus, it became a

[8] https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/

[9] https://towardsdatascience.com/image-captioning-with-keras-teaching-computers-to-describe-pictures-c88a46a311b8

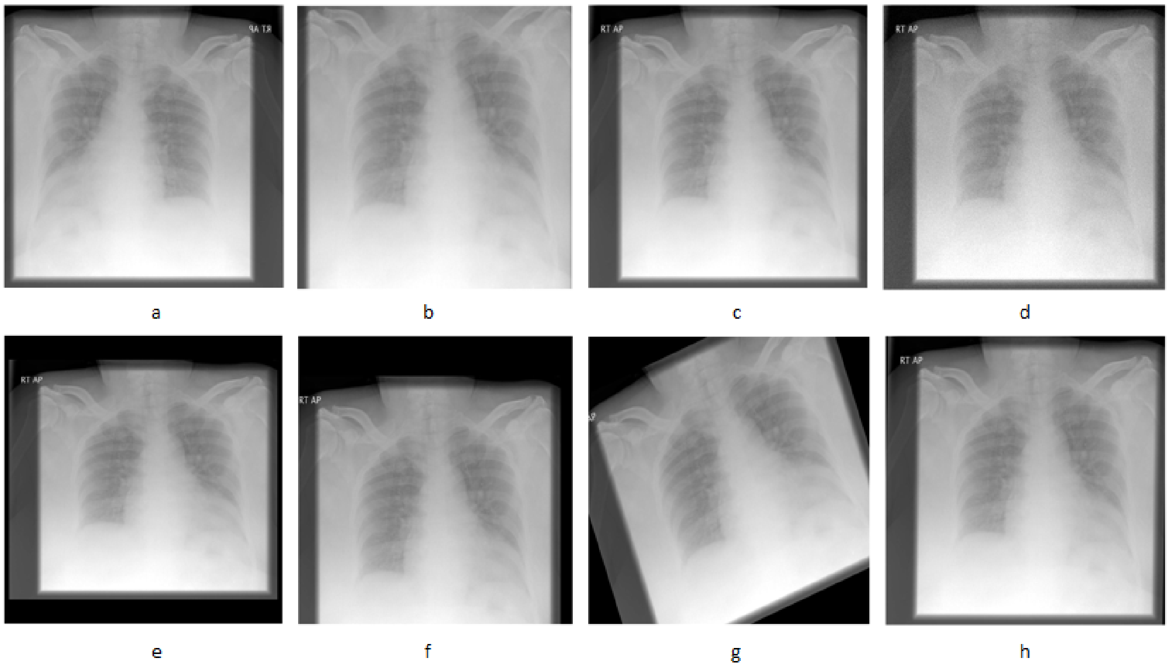


Figure 10: A brief look at augmented images. (a)-flip, (b)-Crop, (c)-Blur, (d)-Noise, (e)-Affine Scale, (f)-Affine Translate, (g)-Rotate, (h)-Shear

necessity to augment these images and train them with more images for better classification results. Image from NIH was not augmented separately because of the large number of images however, OpenI dataset has been augmented with 8 different methods. Figure 10 shows one image which has been augmented in 8 different ways.

3.4 Modelling

Post literature survey, it was seen in the field of face recognition, multi-label classification, video or audio category classification convolutional neural network performed exceptionally good with high prediction accuracy using less computational power than traditional machine learning algorithms. Thus, in our this research, to detect pulmonary fibrosis along with other diseases, if exists, CNN is been used. On the other hand, for multi-annotation and to generate captions for chest x-rays, long short term memory RNN is been utilised. By bringing both the deep learning technique together we are able to multi-classify pulmonary disease and are able to generate a medical report as well. Following are the functions that play an important role in the overall execution of the model:

- **Loss Function:** Loss function captures the inconsistencies between the predicted and actual values and generates an error score.

- **Optimisation Function:** This function helps in reducing the error generated by loss function by backpropagating the weights and make the model optimised.

- **Activation Function:** This helps in deciding when to activate the perceptrons which transfer the information from one later to another with the deep learning model architecture. This is usually placed either in between the layers or at the end of it.

3.4.1 Convolutional Neural Network (ConvNet/CNN)

A combination of multiple convolutional with filters(kernels), pooling, fully connected and an output layer, CNN is majorly used in classification tasks. It is termed as a feed-forward neural network which commonly takes images as input and by using activation, loss and optimisation function converts it into a matrix and transfer it to the next layer. Fully connected layers transform these matrices into a vector of layers which is then used by the output layer to classify the input. The output layer contains perceptrons (neurons) which is equal to the number of classes in the problem. The result is in the form of a percentage, a perceptron outlines the highest percentage is the resulted class.
One such type of CNN is **Streaming CNN**. There are multiple ways of building a streaming CNN. One way that is been followed in this research is transferring weights generated by one CNN into another for better classification results.

3.4.2 Recurrent Neural Network (RNN)

Based on the concept of CNN, RNN is an algorithm that doesn't follow a feed-forward approach but works both ways. RNN can store the information that is processed recently remembers it and produces a sequential result. Apart from the input, hidden and output layer, it also contains a memory that stores information in it and assists the updating the weights with the help of backpropagation. Due to such reasons, the algorithm is most famous in areas like audio translation, time-series analysis, captioning, etc. RNN is very prone to overfitting which is it's one of the weaknesses. One such RNN technique is long short term memory (LSTM).
LSTM: It holds a big memory space and can store more information for a longer time. According to Hochreiter & Schmidhuber (1997), LSTM is capable of performing all DML operation which RNN cannot.

3.5 Evaluation and Deployment

After the modelling, it is now necessary to analyse the performance of our design. Based on data and the output generated, two evaluation techniques have been used. AUROC curve for evaluating the accuracy and prediction power of the CNN and BLEU to compare the score between the predicted and actual caption. Both of these techniques are widely used in their respective areas. The complete process of it is discussed in section 5.

4 Implementation

All the pre-processing, data preparation and modelling is done in Jupyter notebook which is a platform written in Python and hosted on Anaconda. Jupyter notebook provides a web interface to write, edit, delete and modify the code. The whole modelling is performed on TensorFlow framework which helps in performing numeric calculations. Keras package is installed which is essential to run deep learning algorithms. Keras runs on top of TensorFlow and helps in the experimentation of deep learning. Apart from this, packages like panda, numpy, pickle, matplotlib.pyplot are used for successful implementation. When dealing with big data, python is considered to be the most prominent scripting language. It is easy to make libraries which can be loaded later without writing a complete code again.

4.1 Data Preparation and Pre-processing

Below are the steps followed until the actual processing of CNN and LSTM-RNN:

- After setting up the environment all the necessary packages were installed in Anaconda prompt.

- An excel files (Data_Entry) containing the labels for the chest x-ray from NIH is checked for missing values and NAs in R Studio using language R. Later, the images aligned with NAs were deleted manually because there were only 7 in number.

- Dataset is designed in a multi-label fashion which has 836 unique combinations of diseases (For eg. label for a single chest x-ray is; Consolidation, Effusion, Infiltration, Pneumonia. This particular x-ray have multiple problems with Consolidation being the major). Similarly, there is 836 unique pattern of disease which becomes next to impossible to process. Thus, these labels are converted to 14 major diseases.

- Later, the dataset is simplified to a binary form by performing one-hot encoding using lambda function in python.

- Reports downloaded from OpenI are in XML files containing loads of information, therefore, to make the model understand and ease our work, these XML files are converted to an excel format and all the necessary information are fetched in columns. This is done by importing 'etree' library from 'lxml' package with the help of a function.

- These reports were anonymised by replacing the personal information with 'xxxx'. This is then eliminated in language R.

- Augmenting Chest x-rays by importing ImageDataGenerator pre-defined function of its preprocessing.image library. This helps in augmenting the image and directly pass it to the model without saving it into the drive. Shown in figure 6.

- Performing word embedding: The whole process of word embedding is explained in section 3.3. This is performed with the help of Embedding, Tokenizer and pad_sequence functions from keras.layers, keras.preprocessing.text, keras.preprocessing .sequence packages respectively.

- OpenI repository provides about 7K images which are a decent number but to make our model perfect we have to prepare it for any possible noise. Thus, with the help of 'imgaug' package which runs on panda with the help of imageio function. With the help of this function, all these 7K images are converted to 70000 images. This makes the dataset huge again. The parameters are shown in figure 10. The total image is 115,200 which are then fed into the models in different ways as below:

4.2 Multi-label classification model

Based on previous researches, it was concluded that for binary or multi-label classification, CNNs are the best performing algorithms. Many convolutional networks are already trained on millions of images and then their weights are deployed in the form of a library under Keras package. These networks are very easy to call and use. Thus, many researchers leverage the use of these pre-trained nets (VGG16, ImageNet, InceptionV, ResNet, Xception etc) which not only helps in reducing the computational time but also helps in achieving good classification results. One such architecture is MobileNet (Howard et al. (2017)). MobileNet is the best suited CNN for mobile vision or embedded images and also in researches where there is a lack of computational power. None of the experimenters have made use of MobileNet in medical imaging. Medical images are also densely packed and are embedded in such a way that it becomes difficult for CNN to exactly understand the pattern and classify an image with great accuracy.

Our model is divided into two parts; First part of the model is to predict 14 various pulmonary diseases whereas the second part also has a CNN which is directly connected to LSTM for image captioning. The objective of this separation is to design a feature vector that understands every possible lung problem and helps in generating the caption for it.

In the first part, two different CNNs are used. One is MobileNet and another is manually designed and then trained from scratch. A complete MobileNet architecture has 28 layers which are distributed among input, hidden, pooling and fully connected layers. As shown in figure 11, to accomplish our task, MobileNet is extended by adding a GlobalAveragePooling, two dropouts, and two dense layers. These have been added by replacing the outermost layer of MobileNet. Dropout value is kept to 0.5 which protects the model from overfitting. It is used by importing MobileNet library from keras.applications.mobilenet. On the other hand, a 6 layered CNN with 1 fully connected, 5 hidden and 5 pooling layers is designed for a similar task. The intention of creating two CNNs is to compare which of these can fulfill the task with better performance and accuracy. This will help future researchers to understand what needs to be done while conducting similar research. Manually prepared CNN can be seen in figure 12. Both these networks take about 48000 chest x-ray images for training and 5000 for testing from the NIH database and generate a weight that can directly be loaded in the second part of the complete architecture. Sigmoid activation, Adam optimizer and binary_CrossEntropy loss function have been used in both the experiments.

In the second part, a pre-trained CNN; VGG16 is considered from keras.applications.vgg16 library. VGG16 is an award-winning CNN developed by a group from Oxford. This CNN is trained on over 15 million images with their caption. Thus, to leverage the weights of captions to generate our reports, VGG16 has been chosen. The architecture of VGG16 consists of 11 layers; 8 hidden and 3 fully connected with a softmax layer in the end. ReLU activation, categorical_crossentropy loss, and Adam optimizer functions are used for training the CNN. VGG16 which is already upskilled on ImageNet is then again trained with Chest X-rays from OpenI. OpenI contains 7K images which are augmented on 9 different methods and became 70000. Thus, 63000 is used for training and 7K for testing.

4.3 Chest X-ray Captioning

An LSTM-RNN model has been implemented for generating chest x-ray reports. Post extracting x-rays features from VGG16 CNN, a word embedding matrix in the form of a vector is created and transferred to the LSTM model. LSTM expects 256 elements of image inputs. Thus, a softmax layer has been implemented at the end of a VGG model that converts a vector of 4096 elements into 256 elements. LSTM has also been implemented with a 256 memory unit that holds the previous input and matches with the upcoming one to generate output. It takes an input of size 256 and generates an output of size 7579. This RNN model is then fitted with the weights generated by VGG and loaded in the local directory with the help of the model.fit function. The output of this layer gives the name of the disease with its described as given in the datasets.

4.4 Final consolidated model

- **Model 1:** This model is trained twice, first by real chest x-ray images and later with the transformed ones. The objective is to know which type of data suits the most. The respective images are transferred to pre-trained CNN 'MobileNet' which has converted the images into a feature vector. The weight of this model is loaded in the local machine and later move it to another CNN 'VGG16'. Before carrying the weights forward, rigorous performance testing was performed to check the prediction capabilities of the first part. VGG16 takes augmented images as an input, loads the weight created by MobileNet, and finally generates a new feature vector. A new set of weights are created by VGG16 CNN which is again being loaded by LSTM to finally produce artificial captions. LSTM was also been fed by the word embedded feature vector to learn the sequencing of reports from OpenI. A visual illustration of the model is shown in figure 11.

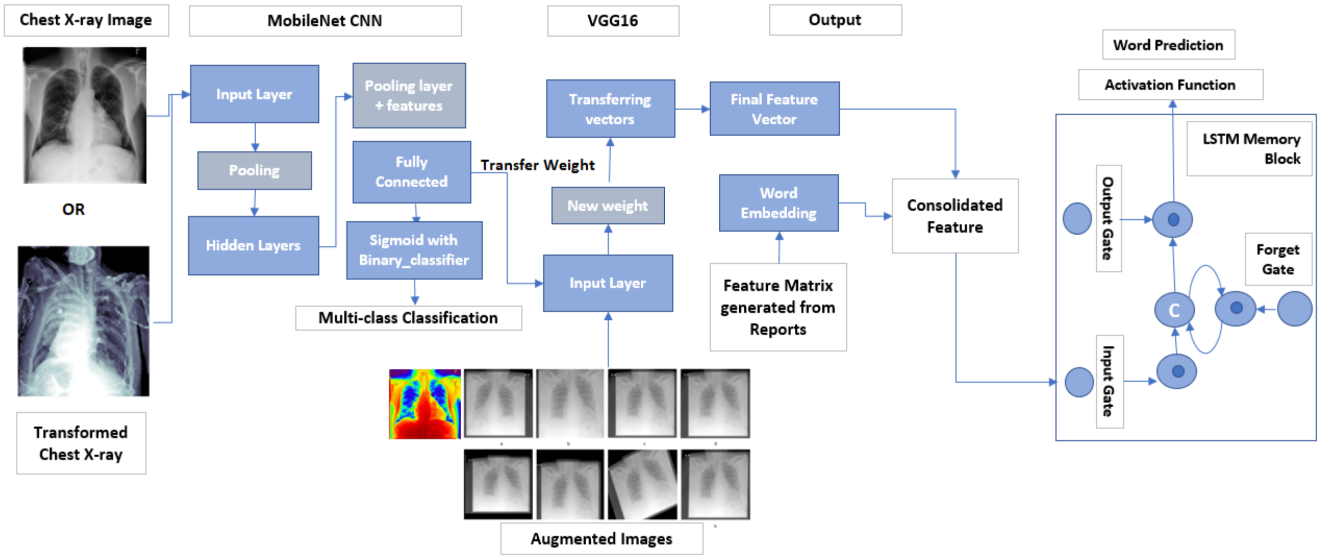


Figure 11: Model 1: Multi-label classification and medical report generation using MobileNet, VGG16, and LSTM-RNN

- **Model 2:** Apart from a slight change, this model is very similar to Model 1. Instead of MobileNet, a self-designed CNN with 5 convolutional layers, 5 pooling, and 1 fully connected layer is used. During the training phase, it was observed that it took comparatively lesser time in completing its execution than Model 1. The complete connectivity is shown in figure 12.

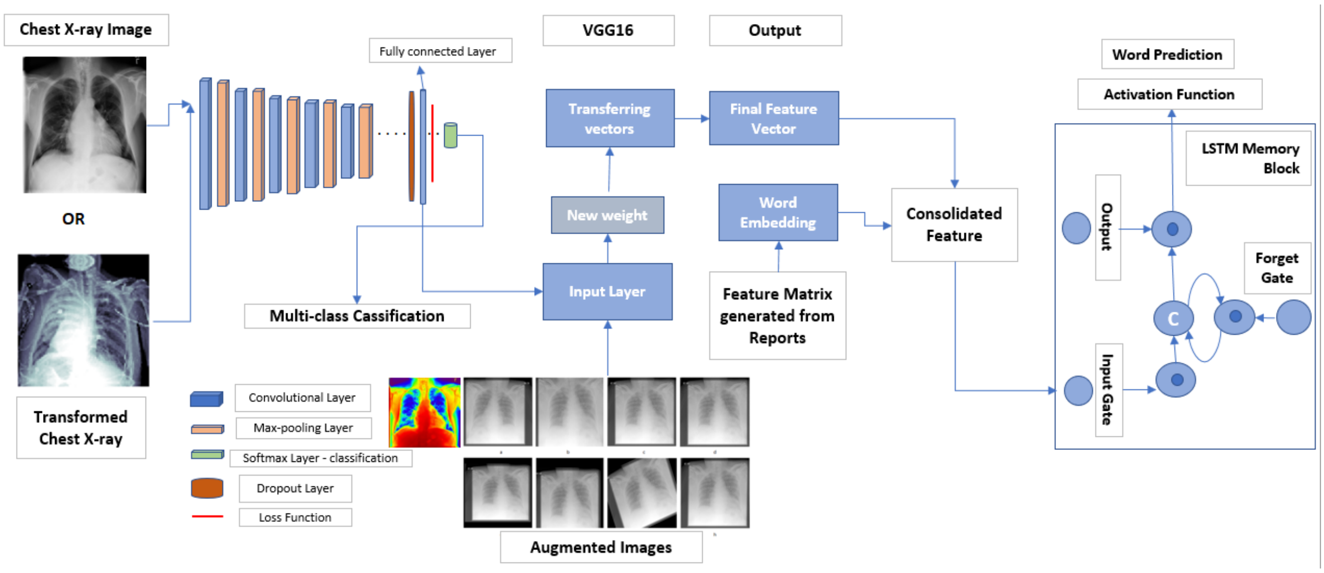


Figure 12: Model 2: Multi-label classification and medical report generation by consolidating a 6 layered CNN , VGG16 CNN, and LSTM-RNN

5 Evaluation

Post successfully implementing the model, it becomes important to evaluate the performance and accuracy. The performance and accuracy of the multi-label classifier are measured with AUROC curve whereas for captioning it is measured with the help of BLEU score. The experiment is performed multiple times by changing the parameters, replacing pre-trained models, augmenting or transforming datasets and also by increasing or decreasing the number of images for training. It was run on 10 epochs with 100 epoch steps each. Detailed discussion on evaluation is as follows:

1. **Area Under the Receiver Operating Characteristics (AUROC):** AUROC has been used in evaluating the performance in most of the researches (Davis & Goadrich (2006)). When dealing with multiple problems, it becomes important to evaluate the model at every step. Thus, as discussed in section 2, AUROC is the best method to evaluate a multi-label classification performance. The curve is shown in figure 13. Also, a table 1 has been prepared to compare the performance accuracy of the multi-label classifier by making the changes in a few variables:

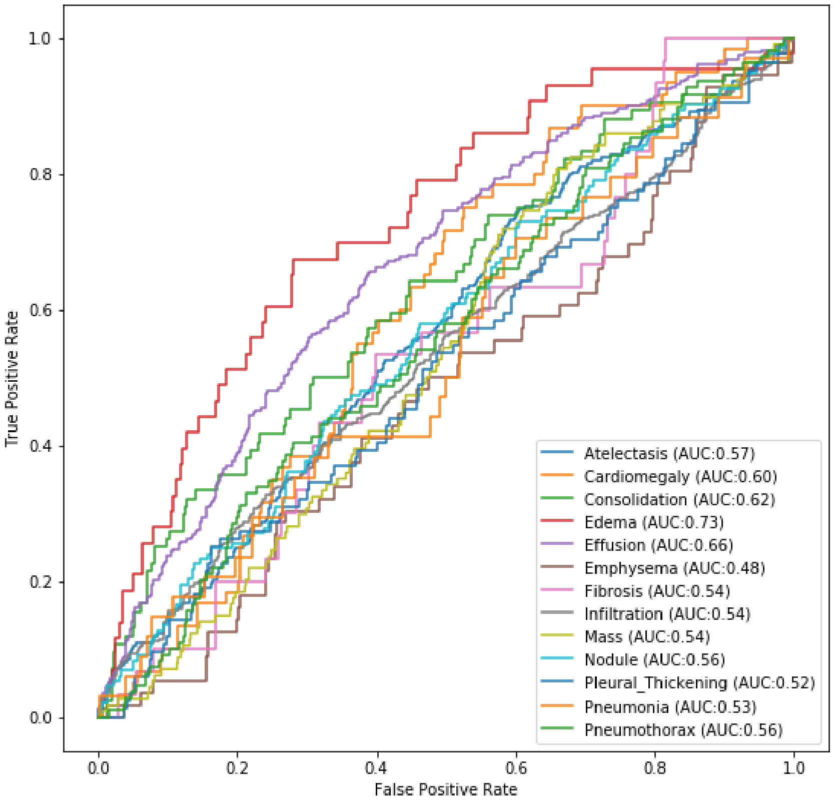


Figure 13: AUROC diagram for multi-label classifier

2. **Bilingual Evaluation Understudy (BLEU):** BLEU is prominently used in machine translation and image captioning (Papineni et al. (2002)). To calculate this score, BLEU compares the predicted caption with the actual one and generates a score. It calculates the n-grams (a sequence of n words taken from the sample) and matches with the prediction. Table 2 shows the BLEU score in some iterations:

Iteration	CNN	Data Type	Activation Function	Accuracy
1	MobileNet	Original Chest X-rays	ReLU	78.24%
2	MobileNet	Original Chest X-rays	Sigmoid	83.12%
3	Self Designed	Original Chest X-rays	ReLU	73.96%
4	Self Designed	Original Chest X-rays	Sigmoid	75.26%
5	MobileNet	Transformed Chest X-rays	ReLU	85.75%
6	MobileNet	Transformed Chest X-rays	Sigmoid	87.57%
7	Self Designed	Transformed Chest X-rays	ReLU	76.95%
8	Self Designed	Transformed Chest X-rays	Sigmoid	77.19%

Table 1: Accuracy of Multi-Label Classifier using MobileNet or Self designed CNN on 48000 images

RNN	Data Type	Training Data Size	Activation Function	BLEU Score
LSTM	Chest X-rays	48000 + 5000	Sigmoid + reLU	0.49
LSTM	Augmented Data	48000 + 63000	Sigmoid + reLU	0.571

Table 2: BLEU scores with MobileNet, VGG16 & LSTM

6 Results & Discussion

Multiple training and testing iterations have been executed to find the best possible combination of parameters to achieve better results in pulmonary disease classification and description generation. The number of images is huge. Hence, to bring the classification capability, the model is trained in parts and the weights are loaded from one to another. In the first part, a multi-label classifier has been generated that predicts the possibility of a disease or a combination of diseases in percentage. Figure 14, shows the output of our classifier when running on some randomly taken test chest X-ray images. Dx signifies the actual labels given in the dataset and PDx signifies the predicted ones. For instance, the first chest x-ray is infected with Effusion and Nodules and prediction also returned the same disease labels with their respective contributed percentage. This reflects the power and capabilities of this multi-class classifier.

The overall diseases detected from test data and predicted outcome percentage is shown in figure 15. For example, it reflects that the overall detection of Cardiomegaly in the test dataset is 5.86% but the model has predicted 5%.

There were 12 iterations in total. Iterations shown in table 1 lists only 8 of them as for the other 4 the model was getting overfit. These 4 executions were performed on unbalanced dataset resulting in “no finding” as the outcome for most of the images.

After successfully preparing this classifier, its weights were loaded in the next part of the complete model. This part makes the use of another classifier i.e. VGG16. It loads the weights generated by MobileNet using transformed images and also takes augmented chest x-ray images gathered from OpenI. Later, the overall weights were transferred to LSTM-RNN which is used as a description generator. The performance of this description generator was measured with the help of BLEU evaluation matrix. The output of one randomly generated chest x-ray is shown in figure 16.

The best accuracy of 87.57% was achieved with the help of MobileNet and Sigmoid activation function using transformed NIH chest x-ray images. When the weights generated by this classifier is loaded by VGG16 in the next part, it generated the BLEU score of 0.57 which is indeed a good one when compared to previous researches performed by Wang, Peng, Lu, Lu & Summers (2018) & Vinyals et al. (2015).

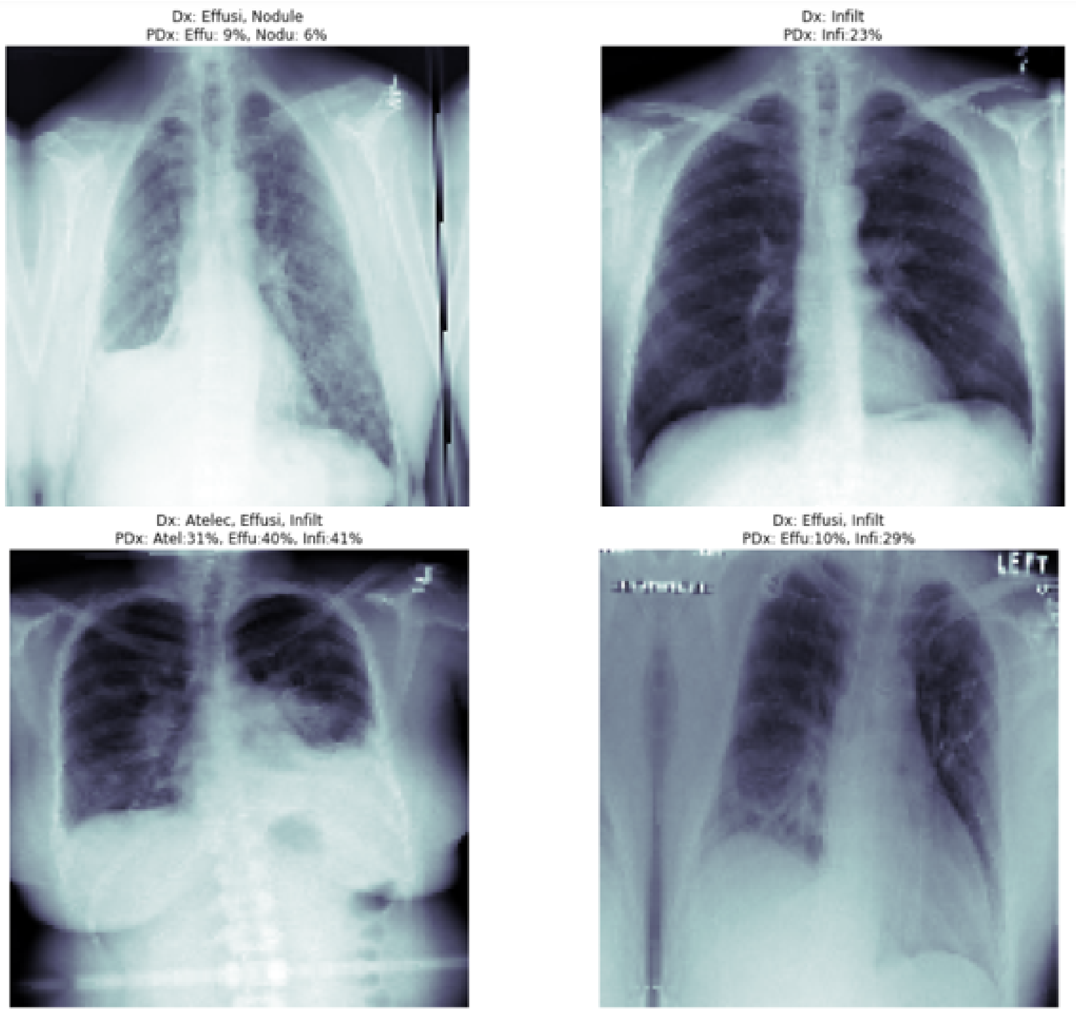
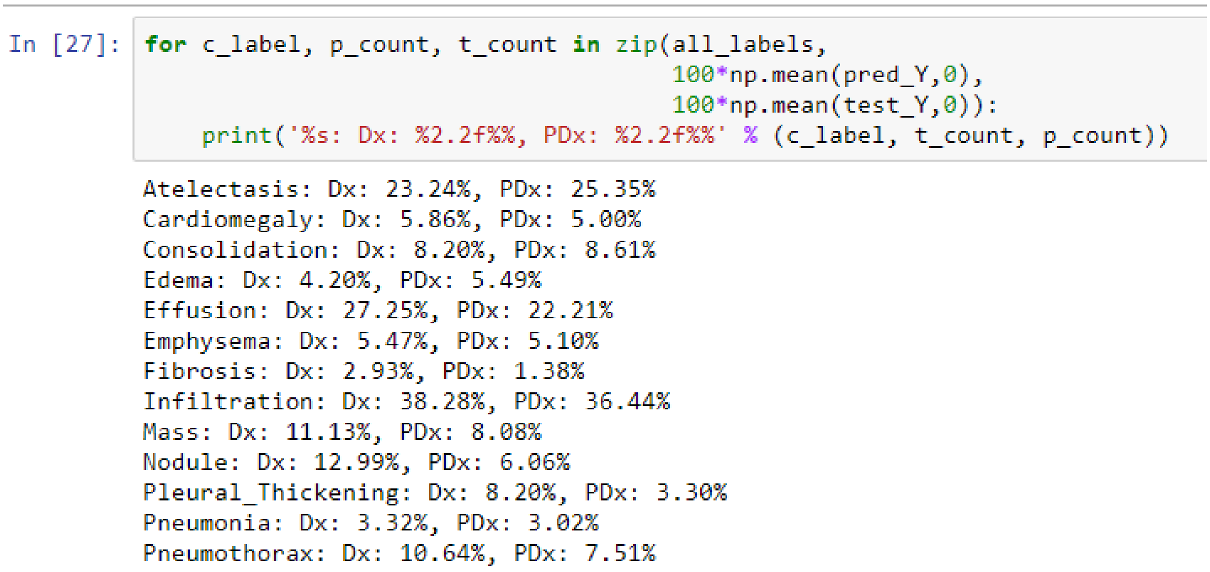


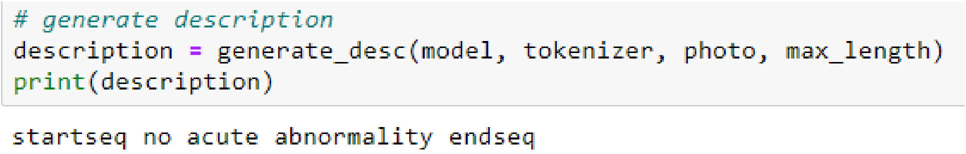
Figure 14: Detected and predicted disease with percentage



```
In [27]: for c_label, p_count, t_count in zip(all_labels,
                                             100*np.mean(pred_Y,0),
                                             100*np.mean(test_Y,0)):
             print('%s: Dx: %2.2f%%, PDx: %2.2f%%' % (c_label, t_count, p_count))

         Atelectasis: Dx: 23.24%, PDx: 25.35%
         Cardiomegaly: Dx: 5.86%, PDx: 5.00%
         Consolidation: Dx: 8.20%, PDx: 8.61%
         Edema: Dx: 4.20%, PDx: 5.49%
         Effusion: Dx: 27.25%, PDx: 22.21%
         Emphysema: Dx: 5.47%, PDx: 5.10%
         Fibrosis: Dx: 2.93%, PDx: 1.38%
         Infiltration: Dx: 38.28%, PDx: 36.44%
         Mass: Dx: 11.13%, PDx: 8.08%
         Nodule: Dx: 12.99%, PDx: 6.06%
         Pleural_Thickening: Dx: 8.20%, PDx: 3.30%
         Pneumonia: Dx: 3.32%, PDx: 3.02%
         Pneumothorax: Dx: 10.64%, PDx: 7.51%
```

Figure 15: Overall detected and predicted disease with percentage



```
# generate description
description = generate_desc(model, tokenizer, photo, max_length)
print(description)

startseq no acute abnormality endseq
```

Figure 16: Sequential output generated by LSTM-RNN

7 Conclusion and Future Work

In this paper, a novel holistic model has been developed to classify pulmonary fibrosis, along with other diseases (if exists) along with generating a medically termed description. The model is divided into two parts; first is working as multi-label classifier whereas the second is working as an image captioner. The objective of creating a classifier is to bring perfection in classifying any pulmonary disease. It is only the weights of the classifier that are transferred to the next part. The second part comprises of a VGG16 CNN and a sequence generator LSTM-RNN. There are two models prepared for the same interest. The only difference is in the classifier. One uses pre-trained MobileNet CNN whereas a self-designed 6 layered CNN is developed for another. It was observed that the computational time of the model with MobileNet was more than the manually created CNN classifier. Prediction accuracy with MobileNet was also comparatively high. Transformation and data augmentation techniques have also played an important role in boosting up the accuracy as well as the BLEU score. AUROC and BLEU evaluation methods have been utilised to measure the performance in parts. Thus, we can conclude that the problem of multi-label classification and description generation of chest x-ray images is solved using this approach. Also, we believe that the model designed is not restricted to work for pulmonary diseases using x-rays but will also work to classify any other disease or object and can also generate a caption for it. It is just that the model needs to be trained on a respective dataset.

In the future, we would like to extend this work by using 3D images, CT scans, and MRIs. 3D images and high-resolution CAD images are a popular research interest Thus, we would like to know how our current methodology performs on such images.

Acknowledgement

I would like to convey my thanks and kind regards to my supervisor Mr. Vladimir Milosavljevic for his supervision and valuable insights. He has been a great motivation and supported me throughout the term. I would also like to thank my friends and family for all their love, support, and encouragement. I am also grateful to NIHCC and OpenI for their exceptionally good and clean datasets.

References

Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A. & Mougiakakou, S. (2016), ‘Lung pattern classification for interstitial lung diseases using a deep convolutional neural network’, *IEEE Transactions on Medical Imaging* **35**(5), 1207–1216.

Attia, M., Hossny, M., Nahavandi, S. & Asadi, H. (2017), Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder, *in* ‘2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)’, IEEE, pp. 3373–3378.

Azevedo, A. I. R. L. & Santos, M. F. (2008), ‘Kdd, semma and crisp-dm: a parallel overview’, *IADS-DM* .

Bai, S. & An, S. (2018), 'A survey on automatic image caption generation', *Neurocomputing* **311**, 291–304.

Cai, J., Lu, L., Xie, Y., Xing, F. & Yang, L. (2017), 'Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function', *arXiv preprint arXiv:1707.04912* .

Chen, G., Ye, D., Xing, Z., Chen, J. & Cambria, E. (2017*a*), Ensemble application of convolutional and recurrent neural networks for multi-label text categorization, *in* '2017 International Joint Conference on Neural Networks (IJCNN)', pp. 2377–2383.

Chen, G., Ye, D., Xing, Z., Chen, J. & Cambria, E. (2017*b*), Ensemble application of convolutional and recurrent neural networks for multi-label text categorization, *in* '2017 International Joint Conference on Neural Networks (IJCNN)', IEEE, pp. 2377–2383.

Davis, J. & Goadrich, M. (2006), The relationship between precision-recall and roc curves, *in* 'Proceedings of the 23rd international conference on Machine learning', ACM, pp. 233–240.

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R. & McDonald, C. J. (2015), 'Preparing a collection of radiology examinations for distribution and retrieval', *Journal of the American Medical Informatics Association* **23**(2), 304–310.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. & Darrell, T. (2015), Long-term recurrent convolutional networks for visual recognition and description, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2625–2634.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. (2017), 'Mobilenets: Efficient convolutional neural networks for mobile vision applications', *arXiv preprint arXiv:1704.04861* .

Jacobs, C., van Rikxoort, E. M., Twellmann, T., Scholten, E. T., de Jong, P. A., Kuhnigk, J.-M., Oudkerk, M., de Koning, H. J., Prokop, M., Schaefer-Prokop, C. et al. (2014), 'Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images', *Medical image analysis* **18**(2), 374–384.

Karpathy, A. & Fei-Fei, L. (2015), Deep visual-semantic alignments for generating image descriptions, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3128–3137.

Kawahara, J. & Hamarneh, G. (2016), Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers, *in* 'International Workshop on Machine Learning in Medical Imaging', Springer, pp. 164–171.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'Advances in neural information processing systems', pp. 1097–1105.

Murphy, K., van Ginneken, B., Schilham, A. M., De Hoop, B., Gietema, H. & Prokop, M. (2009), 'A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification', *Medical image analysis* **13**(5), 757–770.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002), Bleu: a method for automatic evaluation of machine translation, *in* 'Proceedings of the 40th annual meeting on association for computational linguistics', Association for Computational Linguistics, pp. 311–318.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K. et al. (2017), 'Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning', *arXiv preprint arXiv:1711.05225* .

Salehinejad, H., Valaee, S., Dowdell, T., Colak, E. & Barfett, J. (2018), Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks, *in* '2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 990–994.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. & LeCun, Y. (2013), 'Overfeat: Integrated recognition, localization and detection using convolutional networks', *arXiv preprint arXiv:1312.6229* .

Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., Wille, M. M. W., Naqibullah, M., Sánchez, C. I. & van Ginneken, B. (2016), 'Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks', *IEEE transactions on medical imaging* **35**(5), 1160–1169.

Setio, A. A., Jacobs, C., Gelderblom, J. & van Ginneken, B. (2015), 'Automatic detection of large pulmonary solid nodules in thoracic ct images', *Medical physics* **42**(10), 5642–5653.

Shen, W., Zhou, M., Yang, F., Dong, D., Yang, C., Zang, Y. & Tian, J. (2016), Learning from experts: developing transferable deep features for patient-level lung cancer prediction, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 124–131.

Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J. & Summers, R. M. (2016), Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2497–2506.

Tanti, M., Gatt, A. & Camilleri, K. P. (2017), 'What is the role of recurrent neural networks (rnns) in an image caption generator?', *arXiv preprint arXiv:1708.02043* .

Tu, Z., Xie, W., Qin, Q., Poppe, R., Veltkamp, R. C., Li, B. & Yuan, J. (2018), 'Multi-stream cnn: Learning representations based on human-related regions for action recognition', *Pattern Recognition* **79**, 32–43.

Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. (2015), Show and tell: A neural image caption generator, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3156–3164.

Wang, Q., Zheng, Y., Yang, G., Jin, W., Chen, X. & Yin, Y. (2018), 'Multiscale rotation-invariant convolutional neural networks for lung texture classification', *IEEE journal of biomedical and health informatics* **22**(1), 184–195.

Wang, S., Zhou, M., Gevaert, O., Tang, Z., Dong, D., Liu, Z. & Tian, J. (2017), A multi-view deep convolutional neural networks for lung nodule segmentation, *in* '2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)', pp. 1752–1755.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. (2017), Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, *in* '2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)', pp. 3462–3471.

Wang, X., Peng, Y., Lu, L., Lu, Z. & Summers, R. M. (2018), Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, *in* '2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 9049–9058.

Wei, W., Wong, Y., Du, Y., Hu, Y., Kankanhalli, M. & Geng, W. (2017), 'A multi-stream convolutional neural network for semg-based gesture recognition in muscle-computer interface', *Pattern Recognition Letters* .

Wu, Z., Jiang, Y.-G., Wang, X., Ye, H. & Xue, X. (2016), Multi-stream multi-class fusion of deep networks for video classification, *in* 'Proceedings of the 24th ACM international conference on Multimedia', ACM, pp. 791–800.

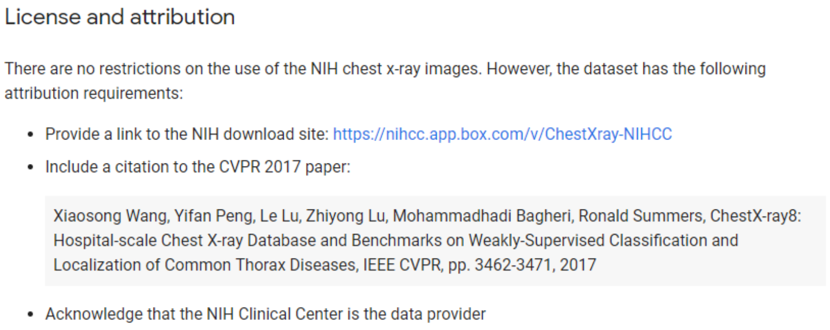
Wu, Z., Jiang, Y.-G., Wang, X., Ye, H., Xue, X. & Wang, J. (2015), 'Fusing multi-stream deep networks for video classification', *arXiv preprint arXiv:1509.06086* .

Xing, L. & Qiao, Y. (2016), Deepwriter: A multi-stream deep cnn for text-independent writer identification, *in* '2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)', IEEE, pp. 584–589.

Xu, X., Guo, Q., Guo, J. & Yi, Z. (2018), 'Deepcxray: Automatically diagnosing diseases on chest x-rays using deep neural networks', *IEEE Access* **6**, 66972–66983.

Appendix

- Licence of NIH dataset:



License and attribution

There are no restrictions on the use of the NIH chest x-ray images. However, the dataset has the following attribution requirements:

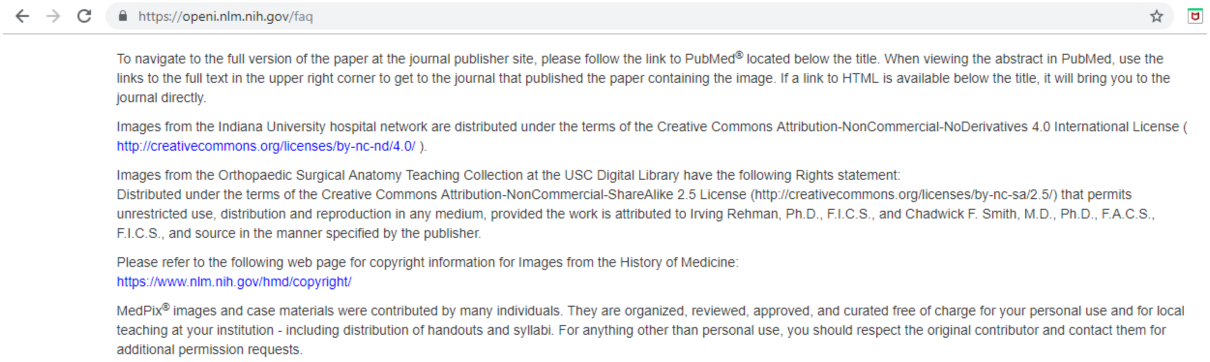
- Provide a link to the NIH download site: https://nihcc.app.box.com/v/ChestXray-NIHCC
- Include a citation to the CVPR 2017 paper:

  Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald Summers, ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462-3471, 2017

- Acknowledge that the NIH Clinical Center is the data provider

Figure 17: Licence of NIH dataset

- Licence of OpenI repository:



To navigate to the full version of the paper at the journal publisher site, please follow the link to PubMed® located below the title. When viewing the abstract in PubMed, use the links to the full text in the upper right corner to get to the journal that published the paper containing the image. If a link to HTML is available below the title, it will bring you to the journal directly.

Images from the Indiana University hospital network are distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Images from the Orthopaedic Surgical Anatomy Teaching Collection at the USC Digital Library have the following Rights statement:
Distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License (http://creativecommons.org/licenses/by-nc-sa/2.5/) that permits unrestricted use, distribution and reproduction in any medium, provided the work is attributed to Irving Rehman, Ph.D., F.I.C.S., and Chadwick F. Smith, M.D., Ph.D., F.A.C.S., F.I.C.S., and source in the manner specified by the publisher.

Please refer to the following web page for copyright information for Images from the History of Medicine:
https://www.nlm.nih.gov/hmd/copyright/

MedPix® images and case materials were contributed by many individuals. They are organized, reviewed, approved, and curated free of charge for your personal use and for local teaching at your institution - including distribution of handouts and syllabi. For anything other than personal use, you should respect the original contributor and contact them for additional permission requests.

Figure 18: Licence of OpenI repository