Lung Cancer Detection

Bharat Venkitesh, Mahesh Kumar Balareddy, and Vignesh Sankar

Electrical and Computer Engineering, University of Waterloo,
Waterloo, Ontario, Canada
{bvenkite,mkbalare,vsankar}@uwaterloo.ca
https://uwaterloo.ca/

Abstract. Lung cancer is the leading cause of cancer-related deaths in US. Early detection and diagnosis gives a good chance for patient survival. In this work our main emphasis on the developing computer aided diagnosis (CAD) tool for early detection of cancer using CT scans as input. We introduce two different data sets, LUNA 16 data set and the kaggle data set. In this paper we present a framework in which we try to predict if a patient has cancer on the kaggle data using information learnt from the LUNA data set. We present the ideas of pre-processing, lung segmentation, lung nodule segmentation and finally classify the data. Owing to the huge size of data, we use GPU enabled clusters to run the algorithms proposed in each step mentioned above. The results obtained were good and the proposed pipeline works and can be used for practical scenarios.

Keywords: Lung Cancer, Convolutional Neural Network, CT scans, UNET

1 Introduction

Lung cancer is the leading cause of cancer-related deaths in the US. In 2012, there were approximately 159,124 related deaths and 229,447 new cases of lung cancer[1]. Early diagnosis is the key for improving the effectiveness of treatment and this potentially increases the patient's chance of survival. Computed tomograpy (CT) and low-dose computed tomography (LDCT) are the most common noninvasive imaging modalities for detecting and diagnosing lung nodules. CT and LDCT can be used for early detection of the nodules as they allow for reconstructing the anatomy of and detecting the anatomic changes in the chest. The goal of this paper is to present a framework that will make use of thousands of high-resolution lung scans provided by the National Cancer Institute to accurately determine when lesions in the lungs are cancerous. This is as part of the Kaggle challenge which we have undertaken. Kaggle is a platform for data science competitions whose main objective is to form strong teams, and use the power of existing data science talent to solve difficult data science problems. Section 2 discusses about the data sets used in detail followed by challenges in section 3. A brief discussion on prior work is in section 4. The proposed framework is introduced in section 5 and the the implementation details along with

results in section 6. We conclude with a few discussion on future work in section 7.

2 Data Introduction

2.1 Kaggle dataset

Kaggle dataset has more than a thousand low-dose CT images from high-risk patients in DICOM format. Each image contains a series with multiple axial slices of the chest cavity. Each image has a variable number of 2D slices, which can vary based on the machine taking the scan and patient. The header of the DICOM file contains the required information about the patient id, as well as scan parameters such as the slice thickness, rescale slope, intercept to name a few. Along with the above data we are provided with the ground truth labels which were confirmed by pathology diagnosis. Data is available at [2]
File Descriptions:
Every patient is identified by patient id and each of them have an associated directory consisting of DICOM files. DICOM header contains the patient id and is similar to the patient name. For each patient we have different number of images which varies according to the number of slices. As the dataset is large, images were compressed as .7z files. DICOM is the standard for medical imaging, which is complex and hence we need to use different available tools to work with these files.

- stage1.7z - All the images to start with the competition. Consists of both the training set and the testing set
- stage1_labels.csv - Ground truth labels for the training set
- data_password.txt - Contains the decryption key for the image files

2.2 LUNA dataset

LUNA dataset is from LIDC/IDRI database which is publicly available. In this dataset we have 888 CT scans. This database also contains annotations collected in two-phase from 4 radiologists. Each radiologists mark the lesions in the lung as non-nodule if the nodule < 3 mm and as nodules if $>= 3$ mm. As part of this dataset we are only considering all nodules which are $>= 3$ mm and which are accepted by 3 out of 4 radiologists. Data is available at [3]
File Descriptions:

- subset0.zip to subset9.zip: Collection of all CT scans.
- annotations.csv: This file contains annotations used for nodule detection.
- candidates_V2.csv: File which contains candidate location to identify false positives.

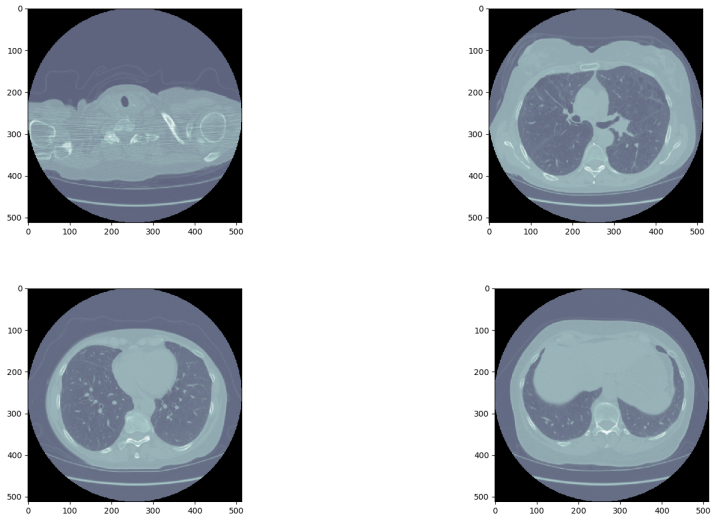

Fig. 2: Lung Segmentation output

3 Challenges

There are several challenges for lung cancer detection and can be summarized as follows.

- The processing is computationally intensive owing to the large dataset (150 GB), hence requires extensive resources to carry out the algorithms.
- The training data (Kaggle data) does not exactly provide us information about which nodules in the lungs as cancerous. (No annotations)
- Irregular size and dimensions of the data. Needs normalization.
- The right threshold has to be chosen so as to accurately separate the lung tissue from the image.
- The nodules are of various densities and are located at disparate locations, so a combination of methods are required to detect them.
- Since this system is used for initial screening, false negatives are given more priority and algorithms need to ensure they reduce the same.
- The dataset is imbalanced, i.e. out of 1384 patients, only 362 patients have cancer. There is a possibility of the classifier over-fitting the data.

4 Literature review

4.1 Review article on Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies.

We reviewed this paper to get an idea on the general framework and the existing methodologies[4].

This paper provides us with schematic diagram of a typical CAD systems for lung cancer detection. The segmentation of lungs in chest images is a pre-processing step which helps in reducing the search space for lung nodules. Followed by detection and segmentation of lung nodules from the reduced search space. Lastly, the classification of the detected nodules is a major component for detection and diagnosis of lung nodules in CT. The resulting CAD system for diagnosis classifies detected nodules into cancerous or non-cancerous.

Lung Segmentation: Existing methods for lung segmentation falls into four categories based on the methods of choice: signal thresholding, deformable boundaries, shape models, or edges. Different methods were used as part of each of the four categories.
Lungs have a high contrast with the surrounding tissue, hence thresholding is the right method to use. As per our needs KMeans with k=2 suits best to find the right threshold. We then use an erosion and dilation techniques to fill in the incursions into the lungs region by radio-opaque tissue, followed by a selection of the regions based on the bounding box sizes of each region to select the region.

Detection of Lung Nodules Simple thresholding, Template matching, Morphological operations, Clustering, LDA's are some of the common techniques used for detecting lung nodules. Feature based classifiers is known to be the best method for classification tasks. Methods like Linear Discriminant Analysis, rule-based or linear classifier, template matching, clustering, neural networks[5] etc are discussed.
We have huge dataset and also have ground truth labels, the frame work we have chosen allows us to use a convolutional neural network to detect and extract nodules.

4.2 U-Net: Convolutional Networks for Biomedical Image Segmentation

The u-net architecture provides very good performance on very varied biomedical segmentation applications[6]. Since it uses data augmentation with elastic deformations, it needs very few annotated images and has a reasonable training time.
U-Net architecture is the method of choice because of the large dataset we are handling, also mainly because we are provided with annotations given as part of the LUNA dataset. Hence this method works well to train and later is used to predict masks on our kaggle dataset.

4.3 Learning Spatiotemporal Features with 3D Convolutional Networks

This paper proposes a very effective approach for spatiotemporal feature learning using deep 3-dimensional convolutional networks trained on a large scale

supervised dataset. As per the findings 3D ConvNets are more suitable for such applications compared to 2D ConvNets and also this produces a good accuracy of 52.8% and is also fast owing to fast inference of ConvNets.
Since we have many 2D slices for each patient. We used this architecture as reference and built our own 3D convolutional neural network architecture for classification.

4.4 Tutorial from Kaggle challenge

Tutorial is available at [7]. With the tutorial as reference for each of the stages discussed, we checked for alternatives and based on the results choose the suitable techniques to get better results.

5 Proposed Framework

A schematic diagram of the framework is given in Figure 3. The basic idea is to leverage the information given from LUNA 16 dataset to predict the nodule locations in the kaggle dataset. Since we use two datasets, a pre-processing step ensures that they are in same field. The segmentation of lung tissues on chest images is an important step to reduce the search space. The two steps are common for both the LUNA and Kaggle datasets. Next, detection and segmentation of lung nodules from the available search space. This is achieved by using LUNA dataset, as it provides us with cancerous regions in the lungs. This is then fed to the kaggle dataset to locate cancerous regions. Finally the classification of the detected nodules into malignant and benign is the final step. The details of each steps are discussed below.

5.1 Pre Processing

As mentioned above, since there are two datasets, one has to make sure that the pixels are in the same range to ensure that the information from one dataset is transferred to the other. To express CT numbers in a standardised and convenient form, Hounsfield unit (HU) is a quantity commonly used in computed tomography (CT) scanning[8]. The kaggle dataset is by default not in this unit. The images are scaled to HU units by multiplying with rescale slope and adding the intercept, that are stored in the metadata of the scans. The LUNA dataset is by default in HU units. The advantage of bringing in to HU units is two fold. Firstly, the datasets are in same range. Secondly, the range of values in HU units represents a physical property like air, lungs, fat, bone etc explained in Table 1. This is used in subsequent steps for lung segmentation.

5.2 Lung Segmentation

The segmentation of lungs in CT images is the second step in the proposed framework for detection of cancer. The segmentation of lungs is a challenging

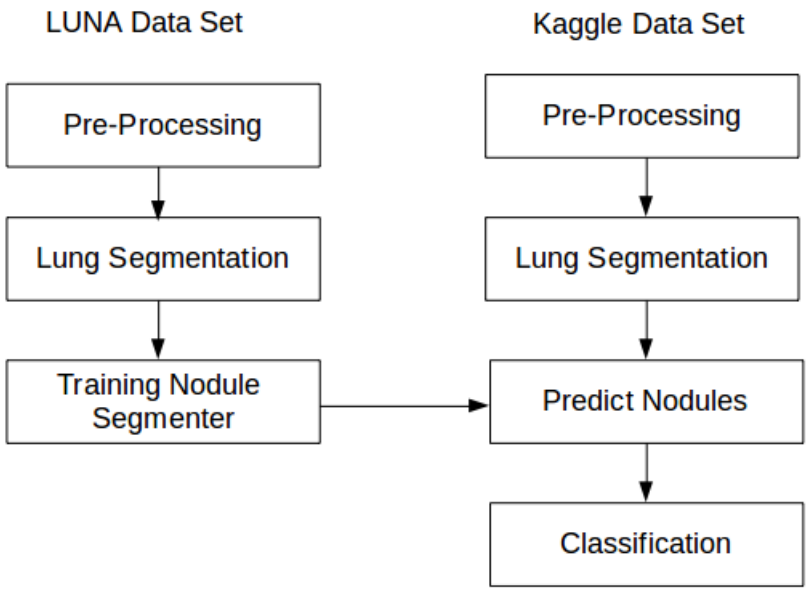

Fig. 3: Proposed Framework

problem due to heterogeneity in lung region and similar densities in pulmonary structures such as arteries, veins, bronchi. The method employed in our framework is to threshold the image to isolate the regions within the image and then separate only the lungs. The threshold is not fixed and can vary across images, this is because certain images have a black background around a grey circular region while others do not as shown in Figure 2. The threshold chosen should be such that it is in between the lung pixel values and denser tissue pixel values. The pixels are reset with the minimum value to the average pixel value of the center of the image(lung region) and perform kmeans clustering to get two clusters. Morphological operation erosion and dilation are applied on the binary image. The methods given above has worked well for a large range of images for both the datasets as shown in Figure 5.

5.3 Lung Nodule Segmentation

After the search space is reduced, the nodules need to be detected. The nodule detection task is a complicated task. The nodules can be found either inside the lungs or on the walls and are very hard to distinguish from shadows, vessels and ribs. We first extract the cancerous regions in images as a mask from the LUNA dataset. The masks and segmented lung images are then fed as input to a CNN to train the network to detect nodules, Figure 7 illustrates this process. The architecture used for cnn is called UNET [6] shown in Figure 8. Unet is a fully

Substance	HU
Air	1000
Lung	500
Fat	100 to 50
Water	0
CSF	+15
Kidney	+30
Blood	+30 to +45
Muscle	+10 to +40
Grey matter	+37 to +45
White matter	+20 to +30
Liver	+40 to +60
Soft Tissue, Contrast	+100 to +300
Bone	+700 (cancellous bone) to +3000 (cortical bone)

Table 1: HU for substances

convoluted network which can trained on very few images and provides precise segmentation results [6]. The network can be divided into two parts. The left side consists of contracting path which can be viewed as feature extraction. The right side consists of a expansive path which takes the extracted features and constructs the mask image. The pooling operation on the right side is replaced by upsampling operators which creates the image with the same size as the input image. As a result, When a test image is given to the UNet architecture, it produces a image of same size with a mask in the nodule location. The trained weights are used to detect nodules in kaggle dataset. The results are shown in Figure 10.

5.4 Classification of Nodules

The final step is to classify if the patient has cancer or not. This step is also not so trivial, as the kaggle dataset does not give us any other information like gender, age which can be used for classification, making it a vision based classification task. The nodules are detected for each image for a patient. Since the sizes vary for different patient, the images are sampled to bring it to a constant dimension. We use a 3D CNN for this with inputs as sampled nodule masks and class label as cancer or no cancer. The architecture of the CNN has 4 layers followed by 2 dense layers.

6 Implementation and Results

As mentioned in section 5.1 the transformed images in HU scale are segmented to get the lung region. Images are then resized to eliminate the non lung region around the lung region. This set of images are then used for CNN for segmentation and classification.

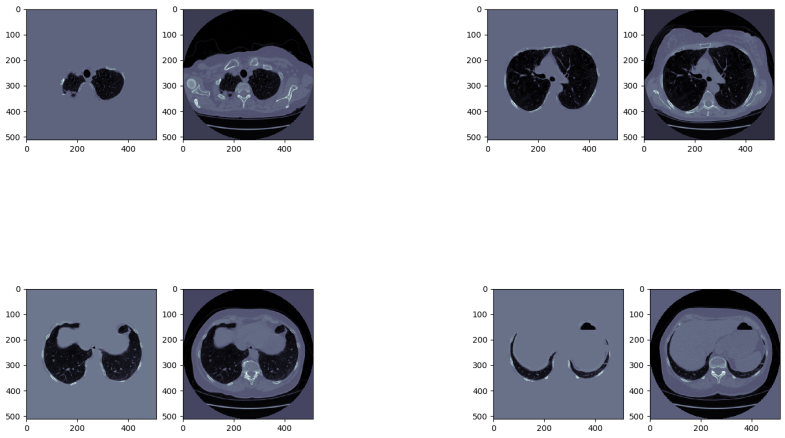

Fig. 5: Lung Segmentation output

6.1 UNET Training

Luna dataset has given annotations or coordinate position and the size of the nodule for cancerous nodules and just the candidate positions for the non cancerous nodules. Keeping the enormous size of the dataset in mind we only consider the cancerous nodule position for training and testing. In order to use the given coordinate information, we need to convert the information given in the world coordinate to the voxel coordinate system. This conversion is done taking the origin position of the CT scan and the spacing between each dimension. The converted coordinates help to find the slice and position of the nodule for a given patient. Because of the 3D nature of the nodules, we take two extra slices from above and below the nodule slice and create 3 masks. As a result one patient on average contributes 3-9 slices to the training and testing dataset. The slices are combined and shuffled together to create the final dataset, totalling 2834 images.

The lung segmented luna dataset is then divided into 3 sets - training, validation and testing sets. 60% of the data is separated and is used for training the unet segmentor. 20% of the data is used as validation set. The result of the validation set is used to fine tune the parameters of the segmentor. The remaining 20% of the Data is used for testing the performance of the segmentor.

We have explored different normalization techniques with the unet architecture to increase the performance of the segmenter. Since we are using various pre-processing algorithms to eliminate unwanted data, we end up significantly changing the dataset and its distribution. So an apt normalization method plays a crucial role for nodule segmentation. Wide range of the HU unit also adds complexity to the normalization procedure. To overcome this challenge we have

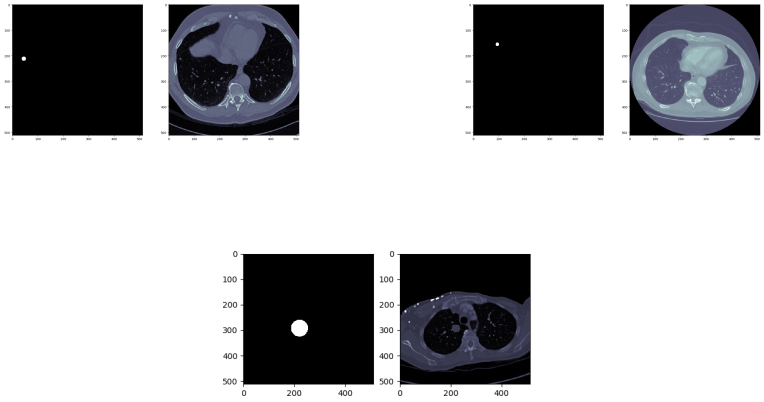

Fig. 7: Nodule mask for images

used various normalization steps at different pre-processing stages and observed the results. The first method used for normalising only the lung regions using the lung mask. The zeros outside the lungs were not considered during normalisation. This method did not give a good performance on the Unet. The second method was to normalise the whole image. Note that in both the above methods, the normalisation is done before resize step. The second method gave us the best result and has been used in our framework.

$$Dice\ coefficient = \frac{2|A.B|}{|A|^2 + |B|^2},\ where\ A\ and\ B\ are\ vectors \tag{1}$$

,

For training purposes, we have used 1500 images and of which 700 images for validation and 600 images for testing the unet architecture. Dice coefficient, given in equation 1 was used as evaluation metric and loss function is the negative of dice coefficient. Initially, there was a big gap between the Training and testing accuracy. To avoid over fitting, we used dropout of 0.2 after every convolutional layer. This increased the performance in the testing dataset. The dice coefficient for each epoch on the validation set is shown in Figure 11, which shows that it increases after each epoch. The fianl values are shown in Table 3. The trained model is used on lung segmented images of the Kaggle dataset to generate the nodule locations.

Dice coreficient	Training set	Testing Set
	0.78	0.53

Table 2: Lung Nodule segmentation results

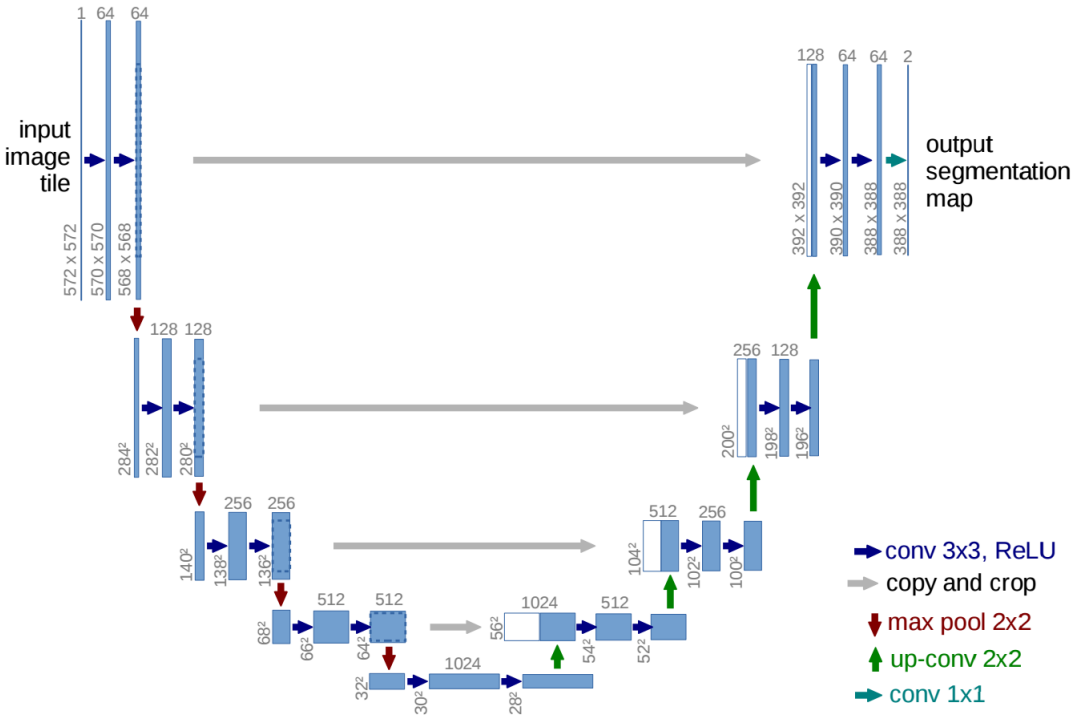

Fig. 8: UNET architecture[6]

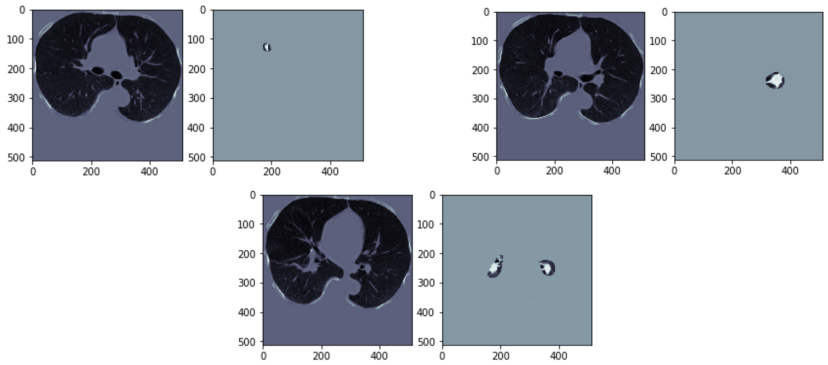

Fig. 10: Predicted Nodules on Kaggle Data

6.2 Nodule Classification

We are using 3D convolutional neural network for classification. Kaggle dataset has 1384 patients and each patient is categorised as cancerous or non cancerous. Since CT scan of each patient is a 3-Dimensional data we choose to use 3D CNN. We have used 3D CNN architecture which is used for learning spatial-temporal features for video classification as reference and created our own 3D architecture [9]. We have followed the same homogeneous architecture where the number of layers is gradually increased to extract the local level features. 4 convolutional layers with increasing number of layers as 32,64,128,256 is used for feature extraction. Max pooling with pool size as (1,2,2)and stride size as (1,2,2) is used between Convolutional layers. 3x3x3 kernel is used for all the layers because of its best performance in video classification. The learned features from the 3D dimensions are flattened and given to the dense layers for classification. We have used 2 dense layers with dropout of 0.5. Relu activation function is used thought out the architecture with the exception of softmax in the final layer. Sparse categorical cross entropy is used as loss metric with Adam optimizer. Number of epochs is 50.

After Nodule segmentation, Adjacent slices are combined to give 20 slices for each patient. The Classifier is trained with 400 patients which makes the training data as 400x20x512x512 dimensional data. This classifier is validated and tested with another 200 patients. With a

6.3 Resources Used

Sharcnet was used for all the computations. It provides a network of high performance computers with the vision to develop and promote high performance computational techniques like deep learning. Sharcnet provides GPU enabled clusters access for 4 hrs at a time with 45GB memory limit. So we can run our deep algorithms in the GPU for 4 hrs continuously. Since training of the Unet and 3D CNN approximately takes 20 mins per epoch. We have saved the weights

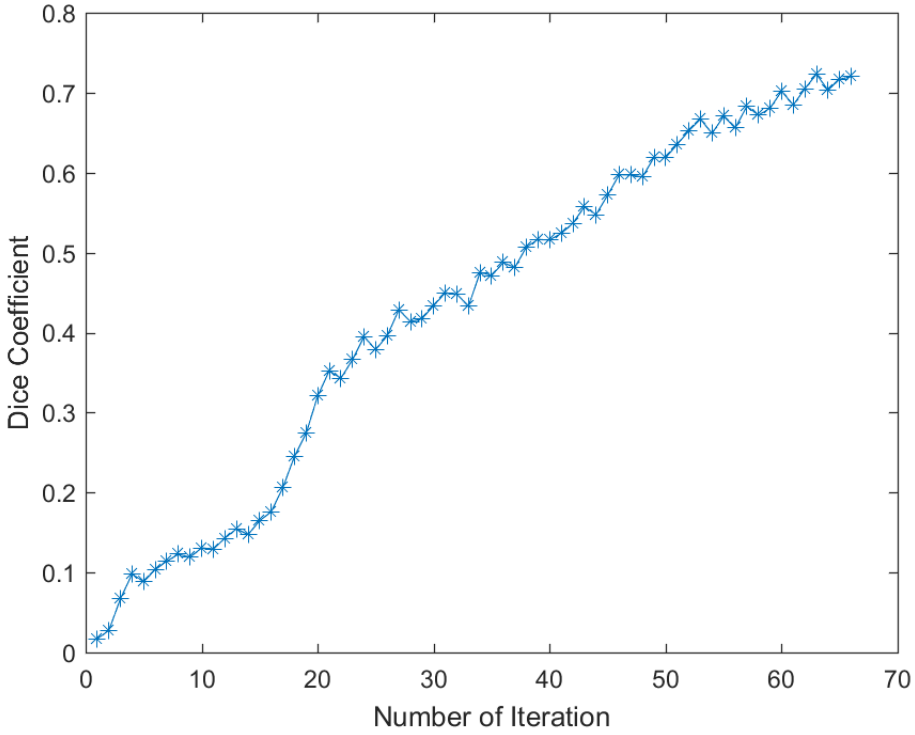

Fig. 11: Dice coefficient against number of epochs

after each epoch and ran the code for multiple days to get sufficient number of epochs.

Accuracy	Training set	Testing Set
	0.67	0.55

Table 3: Lung Classification

7 Conclusion and Future Work

We proposed a general framework for classification of the kaggle dataset by diving the procedure into three categories and using different techniques from the literature for each step. We stated the challenges of the project and discussed how we addressed each one. We covered different aspects of data modelling and analyses and used different algorithms to overcome the challenges at each step. For lung segmentation, we have used image processing techniques and K-means clustering. The nodule segmentation was done using Unet Architecture, where we have explored 2D CNN architecture and Knowledge transfer from one dataset to another.Finally, Nodule Classification was done using 3D CNN where we build a new architecture from the understanding of the Unet architecture and Video classification architecture. In addition, we gained experience in python, sharcnet, managing large dataset and debugging CNN. As future work, Nodule classification can improved by fine tuning the parameters of the 3D CNN or by using a combination of CNN and other feature extraction algorithm.

References

1. American Cancer Society, Cancer facts and figures, 2012.
2. https://www.kaggle.com/c/data-science-bowl-2017/data
3. https://luna16.grand-challenge.org/download/
4. Ayman El-Baz, Garth M. Beache, Georgy Gimel'farb, et al., Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies, *International Journal of Biomedical Imaging*, vol. 2013, Article ID 942353, 46 pages, 2013. doi:10.1155/2013/942353.
5. D. Kumar, A. Wong and D. A. Clausi, "Lung Nodule Classification Using Deep Features in CT Images," 2015 *12th Conference on Computer and Robot Vision*, Halifax, NS, 2015, pp. 133-138. doi: 10.1109/CRV.2015.25
6. O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *MICCAI*, pp. 234241, Springer, 2015.
7. https://www.kaggle.com/c/data-science-bowl-2017tutorial
8. https://en.wikipedia.org/wiki/Hounsfield_scale
9. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri1, "Learning Spatiotemporal Features with 3D Convolutional Networks," *Facebook AI Research, Dartmouth College*