Lung Cancer Detection and Life Expectancy Post Thoracic Surgery Using CNN and Supervised Machine Learning Algorithms

Lavesh S[1], P Sree Lekha[2], Naveen Kumar R[3], Manoj T V[4], Sushila Shidnal[5]

[1,2,3,4]Department of Computer Science Engineering, Sir M Visvesvaraya Institute of Technology, Bangalore, India
[5]Faculty in Dept. of Computer Science Engineering, Sir M Visvesvaraya Institute of Technology, Bangalore, India

Abstract - *Lung Cancer is a disease characterized by the uncontrolled cell growth in tissues of the lungs. It is one of the dangerous and life taking disease in the world. Early detection of lung cancer helps in identification of the treatment to be given which in turn increases the chances of survival of the lung cancer infected patient. This paper presents an approach which utilizes Convolution Neural network (CNN) to classify the CT lung images as cancerous or non-cancerous. The accuracy obtained by means of CNN is 95% which is more efficient when compared to accuracy obtained by the traditional neural network systems. Once the cancer is detected the paper also presents an approach to predict the life expectancy of the lung cancer infected patient post thoracic surgery. The prediction part is implemented using Machine Learning techniques. Linear Discriminant Analysis gave an accuracy of 83.76% with a F-measure of 0.83 which is more compared to other machine learning algorithms used.*

Key Words: **Lung Cancer, Convolution Neural Network, Computed Tomography, DICOM images, Thoracic Surgery, Segmentation, Linear Discriminant Analysis**

1. INTRODUCTION

Cancer is a disease in which cells in the body grow out of control and is one of the most serious health problems in the world. Among various different types of cancer, lung cancer is one of the leading causes of death in both men and women. According to World Health Organization, it was observed that in the year 2018 2.09 million cases of lung cancer was registered and a total of 1.76 million died due to lung cancer. One of the reasons for the high death rate due to lung cancer is the late detection of the lung cancer. Also, the treatment and the prognosis depend on the type of the lung cancer, the stage and the patient's performance. Once the lung cancer is detected, possible treatments include thoracic surgery, chemotherapy and radiotherapy.

At present, there are few biological or technical methods to prevent cancer. Therefore, the project focuses on the early detection of the lung cancer which is the vital factor in the treatment which will in turn increase the survival rate post the treatment. Also, if lung cancer is detected, an attempt is made to predict the survival of the patient for a minimum span of one-year Post Thoracic Surgery which will assist the doctors to take the proper decision for the medication.
The project develops a system that detects the lung cancer by processing the Computed Tomography (CT) image of the lungs under test. The system uses Convolution Neural Network (CNN) which takes CT scan lung images which is present in Digital Imaging and Communications in Medicine (DICOM) format as input and further outputs whether the image is cancerous or non-cancerous. If lung cancer is detected, further the life expectancy post thoracic surgery can be predicted using the Machine Learning techniques. Thus, the system can be taken as an aid for the doctors to effectively make decisions for the treatment of lung cancer patients.

1.1 Literature Survey

Disha Sharma *et al.* (2011), proposed an approach for the early detection of lung cancer by analyzing lungs CT images using Image Processing techniques[1]. The authors used bit-plane slicing, erosion and Weiner filter image processing techniques to extract the lung regions from the CT image. Further the extracted lung regions were segmented using Region growing segmentation algorithm and later Rule based model was used to detect the cancerous nodules. With the help of diagnostics indicator, it was observed that the proposed method achieved an overall accuracy of 80%.

Hamid Bagherieh *et al.* (2013) gave a methodology to detect and classify the lung nodules using Image processing and Decision-Making techniques[2]. Initially, image pre-processing was carried out on a CT images by using contrast enhancement and linear filtering. Next, the filtered image was segmented using Region growing Segmentation process. Further the features like area and color was given as input to the Fuzzy system which employed fuzzy membership function to find the abnormalities.

Sindhu V *et al*. (2014), proposed an approach where the authors aimed to classify the survival of lung cancer patients post thoracic surgery[3]. The authors used Naïve Bayes, PART, J48, OneR, Random Forest and Decision stump algorithm techniques to classify the target function. The performance measurement shows that Random Forest gave high accuracy of 95.65% compared to other ML techniques used.

Prashant Naresh *et al.* (2014), proposed a methodology to detect the lung cancer using Image processing and Neural Techniques[4]. Initially the CT image of lung was filtered to remove Gaussian white noise and Otsu's threshold technique was used to do the segmentation of the image. The structural and features were extracted and these features were given as input to the classifier. The SVM and ANN techniques were used for the classification and it was found that SVM techniques gave a higher accuracy of 95.12%.

Kwetishe Joro Danjuma (2015), proposed a methodology to predict the one-year survival of the patient post thoracic surgery[5]. Naïve Bayes, J48 and Multilayer Perceptron algorithms were used to classify the target class. The Naïve Bayes gave an accuracy of 74.4%, J48 gave an accuracy of 81.8% and MLP gave an accuracy of 82.4%.

Abeer S Desuky *et al.* (2016), proposed a methodology to predict the post-operative life expectancy after Thoracic Surgery using attribute and selection ranking by using Simple Logistic, Multilayer Perceptron and J48 techniques[6]. The Simple Logistic gave an accuracy of 84.68%, MLP gave an accuracy of 81.28% and J48 gave an accuracy of 84.47%.

Peyman Rezaei *et al.* (2017), proposed a methodology to predict the one-year survival of lung cancer patients post thoracic surgery using the combination of Bayesian Network (BN) and Uniform counts discretization[7]. Both BN and Uniform counts discretization yielded an accuracy of 91.28%.

Wafa Alakwaa *et al*. (2017), demonstrated a CAD system for lung cancer classification of CT scans with unmarked nodules[8]. Thresholding was used as initial segment approach which produced best lung segmentation. A modified U-Net trained on LUNA16 data was used to detect nodule candidates. The U-Net output were fed into 3D Convolutional Neural Networks (CNNS) to ultimately classify CT scans as positive or negative for lung cancer. The 3D CNNS produced a test set accuracy of 86.6%.

S Senthil *et al.* (2018), aimed to detect the lung cancer by using neural network with optimal features. Initially the data pre-processing was applied on the input images for the image enhancement[9]. Then the enhanced images were trained and tested by neural network. First, Particle Swarm Optimization (PSO) was applied to extract the features of the input images and the input sample was classified as cancerous or non-cancerous depending on the Artificial Neural Network technique. It was observed that the proposed technique gave an accuracy of 97.8%.

S Sasikala *et al.* (2018), presented an approach to classify the tumors found in lung as malignant or benign[10]. The CT image is pre-processed using median filter technique. Also, back-propagation algorithm was used to train the CNN to detect the lung tumors in CT image. To train the model, lung image with different shape and size of cancerous tissues were fed and the CNN based method was able to detect the presence or absence of cancerous cells with an accuracy of 96%.

2. PROPOSED SYSTEM

A system is developed which detects the lung cancer from the given input CT scanned lung images which are in DICOM (.dcm) format. In addition to this, the system also helps to predict the Life Expectancy Post Thoracic Surgery of the Lung Cancer infected patients.

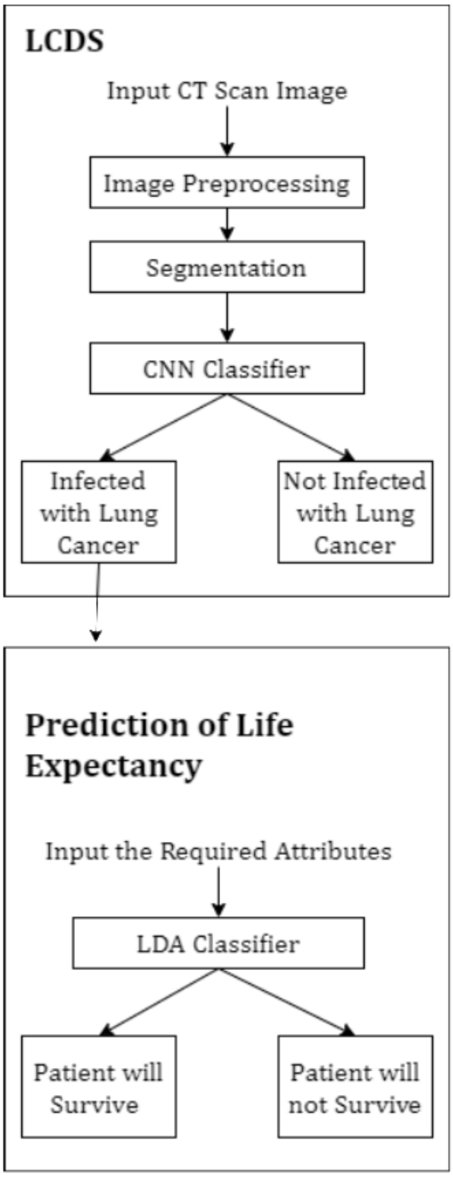


Fig-1: Block Diagram of the System

Fig-1 specifies the block diagram of the system where the entire system is subdivided into two main parts, one is Lung Cancer Detection and the other is the Prediction of Life Expectancy post thoracic surgery.

2.1 Lung Cancer Detection

The CNN model is used to train the CT scanned lung images. Total of 776 images are used to train the network and 90 images are used to validate the network. The Lung Cancer Detection System undergoes the following steps:

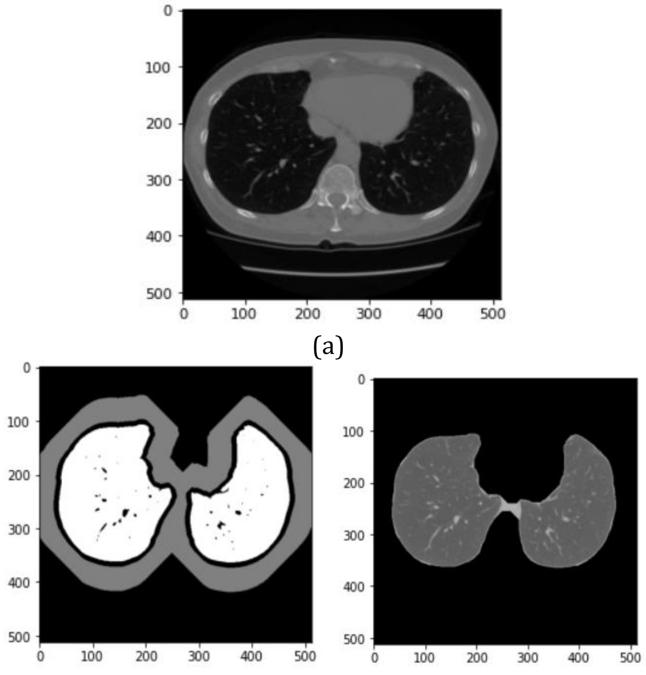
1. Image Preprocessing.
2. Image Segmentation.
3. Training the CNN model.

A. Image Preprocessing

Here, the image originally of size 512 X 512 is filtered and further the image is resized using OpenCV's resize method by using the INTER_CUBIC interpolation.

B. Image Segmentation

Segmentation is the process of partitioning a digital image into multiple segments with the aim to simplify or change the representation of the image into something that is more meaningful and easier to analyze. Here the Watershed Segmentation is used to include the voxels from the edges. Fig-2(a) specifies the original slice of the Lung CT scan image in the gray scale. Fig-2(c) specifies the final segmented lung image produced after applying the watershed segmentation.



(a)

(b) (c)

Fig-2: (a) Original Slice of CT Scan Image (b) Watershed Marked Image (c) Final Segmented Lung Image

C. Training the CNN model

Here, 776 images are used to train the CNN model where each image is either labelled as zero indicating the non-cancerous image or one indicating the cancerous image.

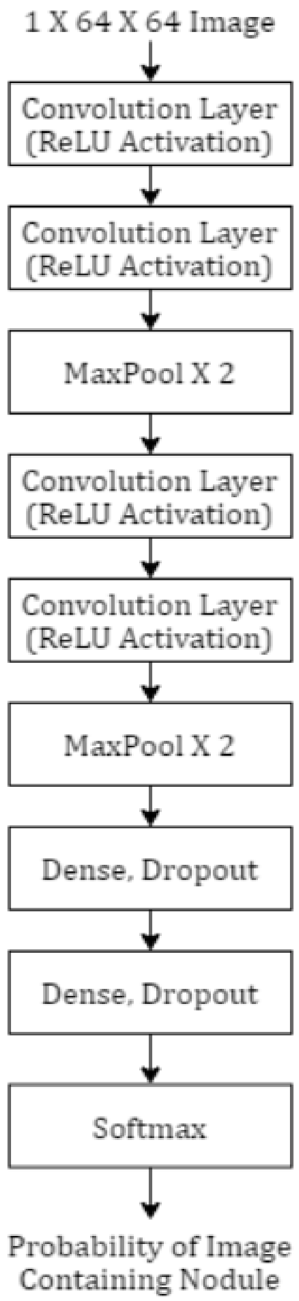

1 X 64 X 64 Image
Convolution Layer (ReLU Activation)
Convolution Layer (ReLU Activation)
MaxPool X 2
Convolution Layer (ReLU Activation)
Convolution Layer (ReLU Activation)
MaxPool X 2
Dense, Dropout
Dense, Dropout
Softmax
Probability of Image Containing Nodule

Fig-3: Architecture of CNN

Fig-3 summarizes the layers used in the Lung Cancer part. CNN also called ConvNet is a class of deep neural networks which is based on shared-weights architecture and translation invariance characteristics. In CNN, the network employs a mathematical operation called Convolution. Convolution is a specialized kind of linear operation. A CNN consists of an input and an output layer with the multiple hidden layers between input and output layer.

The input image is fed into the first convolution layer which uses 4 filters of size 3 X 3 with a ReLU activation. ReLU is the abbreviation of rectified linear unit which applies the non-saturating activation function f(x) = max(0, x). It effectively removes negative values from an activation map by setting them to zero. Further one more convolution layer is used with 4 filters of size 3 X 3 with a ReLU activation. Further pooling is applied which progressively reduce the spatial size of the representation to reduce the number of parameters and computation in the network. Pooling layer operates on each feature map independently. In the LCDS system, Max Pooling is used with a 2 X 2 window and a stride of 2. Next, again a convolution layer with the 8 filters of size 3 X 3 is used with ReLU activation. The above layer is repeated again. Then the Max Pooling function is applied with the same window size and stride as the precious one. Next a dense layer of size 32 is used which acts as a classic fully connected neural network. Next a dropout layer is used to tackle overfitting where the dropout value of 0.5 is used. One more dense and the dropout layer is used with the same parameters. At last a SoftMax loss is used for predicting a single class out of K mutually exclusive classes.

The Sequential CNN model is compiled with a Stochastic Gradient Descent (SGD) optimizer which is an iterative method for optimizing an objective function with suitable smoothness properties. In this system, the learning rate of 1e-2, decay of 1e-6, momentum of 0.9 and nesterov of 'True' value is used as parameters for SGD. Also, the model is compiled with the 'categorial_crossentropy' loss.

2.2 Prediction of Life Expectancy

This is the second part of the system which aims at predicting the survival of lung cancer infected patient post thoracic surgery.

The dataset includes 17 attributes which are specified in Table-1. Among all those attributes, Risk1Y is the target class specifying zero if the patient survives for at least one-year post thoracic surgery and one for those who died before completing one-year post surgery.

The visualization of the dataset is done using the Matplotlib and Seaborn libraries of Python. Further the essential attributes are found based on the Information Gain (IG) attribute evaluation which is used to find the importance of an attribute by using the Information Gain with respect to the target class.

IG (Class, Attribute) = E (Class) – E (Class | Attribute)

where, E stands for Entropy

After the IG Attribute Evaluation on all the 16 independent attributes in the dataset, it is found that the attributes PRE19 and PRE32 gives the IG value as zero and hence are the least useful attributes for training the model. Therefore, these two attributes are eliminated and the remaining 14 attributes are used to train the model.

Table -1: Dataset Attributes

Name	Description	Type
DGN	Diagnosis – specific combination of ICD-10 codes.	Nominal
PRE4	Forced Vital Capacity – FVC	Numeric
PRE5	Forced Expiratory Volume – FEV!	Numeric
PRE6	Performance Status – Zubrod Scale	Nominal
PRE7	Pain before Surgery	Binary
PRE8	Haemoptysis before Surgery	Binary
PRE9	Dyspnoea before surgery	Binary
PRE10	Coughing up blood before surgery	Binary
PRE11	Weakness before Surgery	Binary
PRE14	T in clinical TNM-size of original tumor, From OC11 (smallest) to OC14(largest)	Nominal
PRE17	Type 2 DM – Diabetes Mellitus	Binary
PRE19	MI up to 6 months	Binary
PRE25	PAD – Peripheral Arterial Disease	Binary
PRE30	Smoking	Binary
PRE32	Asthma	Binary
AGE	Age at the time of Surgery	Numeric
RISK1Y	1 Year Survival (T-Died, F- Alive)	Binary

To train the dataset the following algorithms are used:

Random Forest: It's an ensemble training method for classification, regression that operates by constructing a multitude of decision trees at training time and outputting the class that is mode of the classes of the individual trees.

Linear Discriminant Analysis (LDA): It's a dimensionality reduction technique which try to reduce the number of dimensions in the dataset while retaining as much information as possible.

Support Vector Classifier (SVC): It tries to find a hyperplane in an N-dimensional space (N specifies the number of features) that distinctly classifies the data points.

Logistic Regression: It's an algorithm which is used to predict the probability of a target variable.

To train the dataset, the principle of 10-fold Cross Validation is used. Here, at first 10 equal sized datasets are created from the original data. Further each data set is partitioned into mainly two parts, that is 90% for training and remaining 10% for testing. After this, a classifier is produced from 90% labelled data and applied to 10% test data for set 1. The above procedure is repeated for remaining sets that is from 2 to 10.

3. RESULT ANALYSIS

3.1 Lung Cancer Detection

The training and the validation images are run for 20 epochs with the learning rate of e-2.

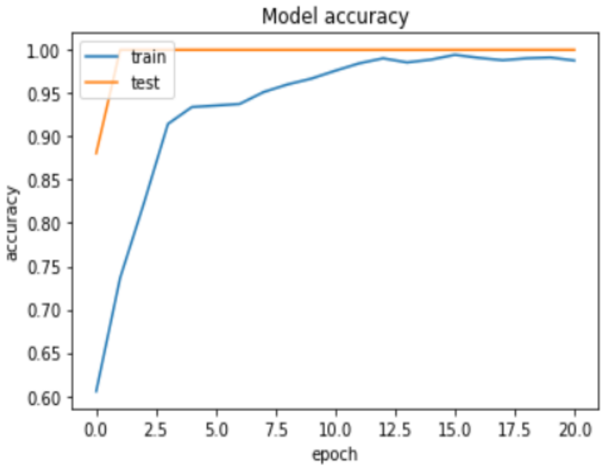


Chart-1: Accuracy of CNN model

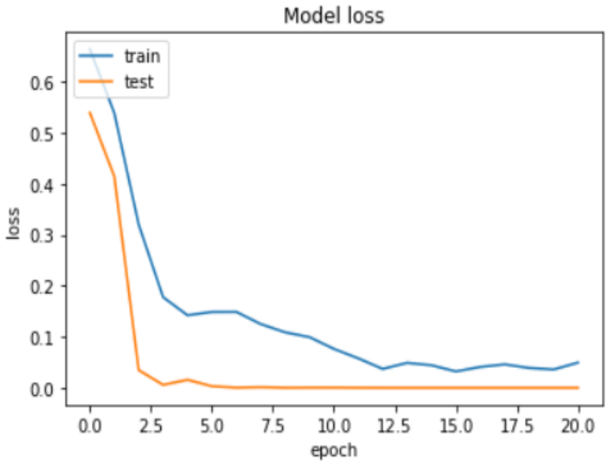


Chart-2: Loss of CNN model

Chart-1 shows the graph of the accuracy with respect to the epochs. It's observed that as the epochs increases, the accuracy of the model also increases. Further Chart-2 shows the loss function and it can be clearly seen that the loss decreases with increase in the epochs.

It is observed that the CNN model gave an accuracy of 95%.

3.2 Prediction of Life Expectancy

The 10-fold Cross Validation is used to train and test the dataset over mainly 4 algorithms that were specified before.

Accuracy alone is not enough to tell which model is the best one and hence along with the accuracy, the F-measure is also calculated to evaluate the model. Accuracy and F-measure are the parameters calculated from the Confusion Matrix representation as shown in Table-2.

Table-2: Confusion Matrix

CONFUSION MATRIX		Predicted Value	
		P	N
Actual Value	P	TP	FN
	N	FP	TN

Table-3: Performance Evaluation of Classifiers

Classifier	Measures			
	Accuracy (%)	Precision (%)	Recall (%)	F-Score
Random Forest	81.80	83.09	81.60	0.82
LDA	83.76	82.59	83.76	0.83
SVC	78.43	81.50	78.43	0.80
Logistic Regression	82.60	80.54	82.61	0.81

Accuracy: It's the percentage of observation that are correctly predicted by the model.

Accuracy = (TP + TN) / (TP + TN + FP + FN)
where,
TP – True Positive
FP – False Positive
TN – True Negative
FN – False Negative

F-Measure: It's also called F-Score which is a harmonic mean of the precision and recall.

F-measure = (2 * Precision * Recall) / (Precision + Recall)
where,
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)

Table-3 specifies the accuracy, precision, recall and F-Measure of the four algorithms used. It's observed that the

Random Forest, LDA, SVC and Logistic Regression gave an accuracy of 82.6%, 83.76%, 78.43% and 81.6% respectively. Therefore, its clear that among the four classifiers used, Linear Discriminant Analysis (LDA) gave the highest accuracy of 83.76% with the F-Measure of 0.82.

4. CONCLUSION

Detection of lung cancer is one of the challenging problems in medical field due to structure of cancer cells, where most of the cells are overlapped to each other. Detection of lung cancer in the early stage is curable. The system contains two parts. One is Lung Cancer Detection part and the other is the Prediction of Life Expectancy Post Thoracic Surgery. Both the parts can run independently. The system is provided as a web application where anyone can upload a CT scan image of the lung in the front end and check out whether that image is infected with Lung Cancer. At the backend the image uploaded is preprocessed, segmented and predicted using the CNN model which is already trained. Also, the system provides a form requesting for the 14 attributes required to predict the survival span post Thoracic Surgery. Once the form is submitted, the form inputs are run on the LDA model and a response of whether the patient will survive or not is produced. CNN model used to detect lung cancer gave an accuracy of 95% and the LDA classifier gave the highest accuracy of 83.76% compared to other 3 algorithms used. The system considers only CT lung images as input, but further the system can be enhanced to take MRI (Magnetic Resonance Imaging) or PET (Position Emission Tomography) as input. Also, in the prediction of survival part, higher accuracies of the classifier can be achieved by considering large amount of dataset.

REFERENCES

[1] Sharma, Disha, and Gagandeep Jindal. “Identifying lung cancer using image processing techniques.” In International Conference on computational Techniques and Artificial intelligence (ICCTAI), vol.17, pp. 872-880. 2011.

[2] Bagherieh, Hamid, Atiyeh Hashemi, and Abdol Hamid Pilevar. "Mass detection in lung CT images using region growing segmentation and decision making based on fuzzy systems." International Journal of Image, Graphics and Signal Processing 6, no. 1 (November 2013):1.

[3] Sindhu, V., S. A. S. Prabha, S. Veni, and M. Hemalatha. "Thoracic surgery analysis using data mining techniques." International Journal of Computer Technology & Applications 5 (2014): 578-586.

[4] Naresh, Prashant, and Dr RajashreeShettar. "Early detection of lung cancer using neural network techniques." Prashant Naresh Int. Journal of Engineering Research and Applications 4, no. 8 (2014): 78-83.

[5] Danjuma, Kwetishe Joro. "Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients." arXiv preprint arXiv:1504.04646 (2015).

[6] Desuky, Abeer S., and Lamiaa M. El Bakrawy. "Improved prediction of post-operative life expectancy after Thoracic Surgery." Advances in Systems Science and Applications 16, no. 2 (2016): 70-80.

[7] Hachesu, Peyman Rezaei, Nazila Moftian, Mahsa Dehghani, and Taha Samad Soltani. "Analyzing a Lung Cancer Patient Dataset with the Focus on Predicting Survival Rate One Year after Thoracic Surgery." Asian Pacific journal of cancer prevention: APJCP 18, no. 6 (2017): 1531.

[8] Wafaa Alakwaa, Mohammad Nassef, Amr Badr. “Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)”. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 8, 2017

[9] Senthil, S., and B. Ayshwarya. "Lung cancer prediction using feed forward back propagation neural networks with optimal features." International Journal of Applied Engineering Research 13, no. 1 (2018): 318-325.

[10] Sasikala, S., M. Bharathi, and B. R. Sowmiya. "Lung Cancer Detection and Classification Using Deep CNN." International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-2S December, 2018