International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 6, pp 222-229, June-2022

https://doi.org/10.47001/IRJIET/2022.606034

Lung Cancer Detection Using Machine Learning

[1]Anusha S, [2]Chandrakumar, [3]Gunashree K, [4]Suma N, [5]Dr. Madhu B K

[1-4]UG Student, Department of Computer Science & Engineering, Vidya Vikas Institute of Engineering and Technology, Mysore, India

[5]Professor, Department of Computer Science & Engineering, Vidya Vikas Institute of Engineering and Technology, Mysore, India

***Abstract* - Cancer is a disease in which cells in the body grow out of control. When cancer starts in the lungs, it is called lung cancer. Lung cancer begins in the lungs and may spread to lymph nodes or other organs in the body, such as the brain. Cancer from other organs also may spread to the lungs. When cancer cells spread from one organ to another, they are called metastases. Lung cancers usually are grouped into two main types called small cell and non-small cell (including adenocarcinoma and squamous cell carcinoma). These types of lung cancer grow differently and are treated differently. Non-small cell lung cancer is more common than small cell lung cancer. Our project consists of development of a machine learning algorithm to detect cancer using x-ray image.**

Keywords: lung disease, cancer detection, machine learning, CNN.

I. INTRODUCTION

Cancer is a fatal illness often caused by genetic disorder aggregation and a variety of pathological changes. Cancerous cells are abnormal areas often growing in any part of human body that are life- threatening. Cancer also known as tumor must be quickly and correctly detected in the initial stage to identify what might be beneficial for its cure. Even though modality has different considerations, such as complicated history, improper diagnostics and treatment that is main causes of deaths.

A thin lining layer called the pleura surrounds the lungs. The pleura protect our lungs and help them slide back and forth against the chest wall as they expand and contract during breathing. Below the lungs, a thin, dome-shaped muscle called the diaphragm separates the chest from the abdomen. When you breathe, the diaphragm moves up and down, forcing air in and out of the lungs.

The aim of is to analyze, review, cancer detection using machine learning techniques for lung cancer. The study highlights how cancer diagnosis, cure process is assisted using machine learning with supervised, unsupervised and deep learning techniques. Several states of art techniques are categorized under the same cluster and results are compared on benchmark datasets from accuracy, sensitivity, specificity, false-positive metrics. Finally, challenges are also highlighted for possible future work.

II. LITERATURE SURVEY

Research and Development on cancer detection is more on imaging than textual data. With the help of documented symptoms in the form of text and Machine Learning (ML) techniques, it is possible to predict the lung cancer stages effectively. This paper conjectures the oeuvre model which is efficient in predicting the stages of lung carcinoma by applying the concepts of ML algorithms. The proposed model is combination of K-Nearest Neighbours, Decision Tree and Neural Networks models along with bagging ensemble method for enhancing the accuracy of the overall prediction. The predicted results of the suggested model are showing better accuracy compared to individual algorithms.

Machine learning techniques are being used in cancer research for more than a decade. Nowadays, Machine Learning Algorithms (MLA) can contribute significantly to the area of Lung cancer (LC) research. LC accounts for the highest mortality rate across the globe, hence early prediction and classification of cancer cells can increase the survival rate substantially. Though there are many algorithms used in the field of neurology, radiology, oncology for LC prediction, ML outperforms those algorithms due to their accuracy and efficiency. This study first focuses on the workflow methodology used by ML for early prediction and classification of LC.

Machine learning (ML) is a significant subset of Artificial Intelligence (AI) that plays a key role in medical diagnosis. The advantage of AI is they can automatically learn, extract and translate the features from data sets such as images, text or video, without introducing traditional hand-coded code or rules. This paper focuses on recognizing and classifying lung diseases by ML algorithms. It includes 400 lung disease images (i.e. CT scan images) including bronchitis, emphysema, pleural effusion, cancer, and normal. The input image is analyzed, categorized and classified using ML algorithms such as the MLP, KNN and SVM classifier. After feature extraction, the output is segmented and compares the classifier's accuracy. When a CT scan image was given to a classifier as an input, it contains irrelevant information. For

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 6, Issue 6, pp 222-229, June-2022
https://doi.org/10.47001/IRJIET/2022.606034

the selection of the most relevant features (i.e. for extracting characteristics) here Gray Level Co-occurrence Matrix (GLCM) is used. For MLP, this classifier acquires 98% accuracy, for SVM accuracy is 70.45% and for KNN accuracy is 99.2%. These classifiers will help the doctors to prescribe the most effective treatment for a patient.

III. ARCHITECTURAL DESIGN

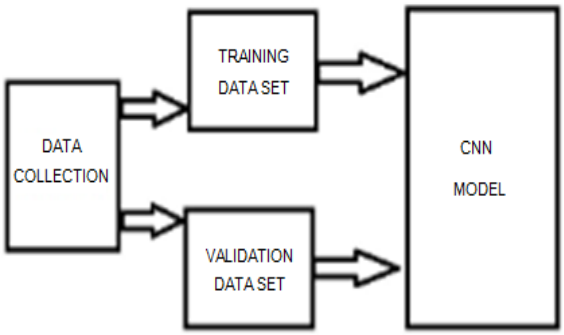

Figure 1: Architectural Design

The system consist of 2 parts

1. Training of machine learning module with data sets
2. Diagnosis of cancer

Training of machine learning module with data sets: The neural machine learning module is created using keras library and python in Google colab editor for training the module 1000's of images of cancerous and non cancerous sample are collected and stored in a file. The file is divided into two parts training datasets and testing data sets training data sets is used to train the machine learning module .It consists of cancerous and non cancerous images labeled respectively. These images will be fed into machine learning module, and trained module is extracted.

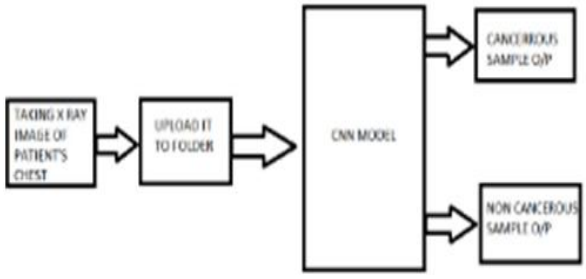

Figure 2: Training of machine learning module with data sets

3.1 Diagnosis of cancer

The diagnosis of cancer starts with the collection chest x-ray at the radiology, the x-ray image is then passed into trained neural module for diagnosis. The trained machine learning module will give output whether the sample is cancerous or non cancerous based on prediction value.

3.2 Flow Chart of the System

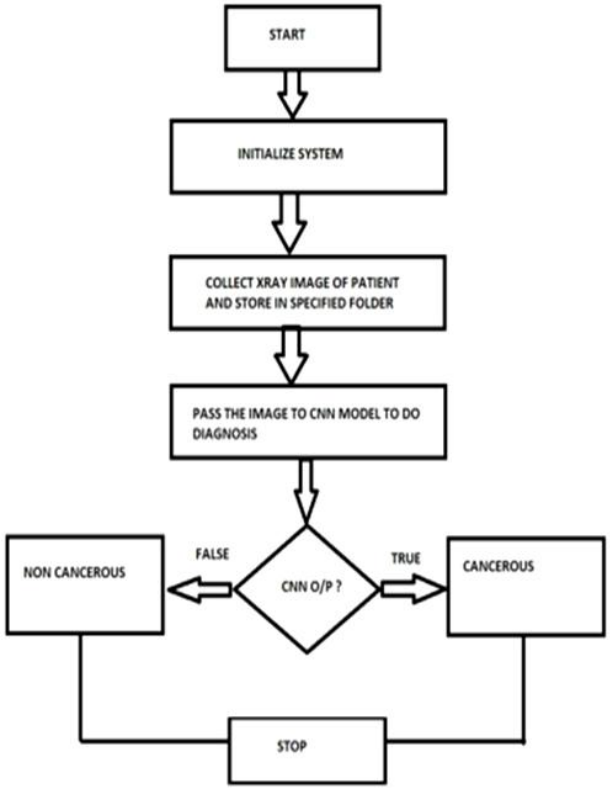

Figure 3: Flow Chart of the System

IV. IMPLEMENTATION

4.1 Overview

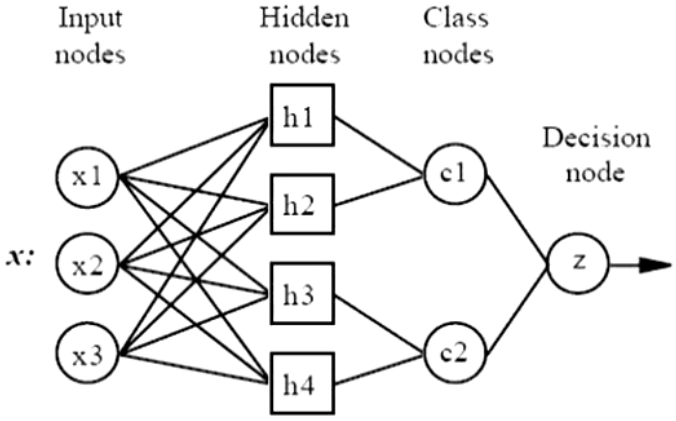

Figure 4: Architecture of convolution neural network

This chapter deals with the implementation process of the project. This chapter covers the data mining process and the software development process. The tasks of the project have been broken down into the following work flow diagram. The Data Understanding and Data Preparation steps have been inspired and adapted from Kaggle by Mulholland et al. and Zuidhof.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 6, pp 222-229, June-2022

https://doi.org/10.47001/IRJIET/2022.606034

4.2 Layers

PNN is often used in classification problems. When an input is present, the first layer computes the distance from the input vector to the training input vectors. This produces a vector where its elements indicate how close the input is to the training input. The second layer sums the contribution for each class of inputs and produces its net output as a vector of probabilities. Finally, a complete transfer function on the output of the second layer picks the maximum of these probabilities, and produces a 1 (positive identification) for that class and a 0 (negative identification) for non-targeted classes.

4.3 Input Layer

Each neuron in the input layer represents a predictor variable. In categorical variables, N-1 neurons are used when there are N numbers of categories. It standardizes the range of the values by subtracting the median and dividing by the inter quartile range. Then the input neurons feed the values to each of the neurons in the hidden layer

4.4 Pattern Layer

This layer contains one neuron for each case in the training data set. It stores the values of the predictor variables for the case along with the target value. A hidden neuron computes the Euclidean distance of the test case from the neuron's center point and then applies the radial basis function kernel function using the sigma values.

4.5 Summation Layer

For PNN networks there is one pattern neuron for each category of the target variable. The actual target category of each training case is stored with each hidden neuron; the weighted value coming out of a hidden neuron is fed only to the pattern neuron that corresponds to the hidden neuron's category. The pattern neurons add the values for the class they represent.

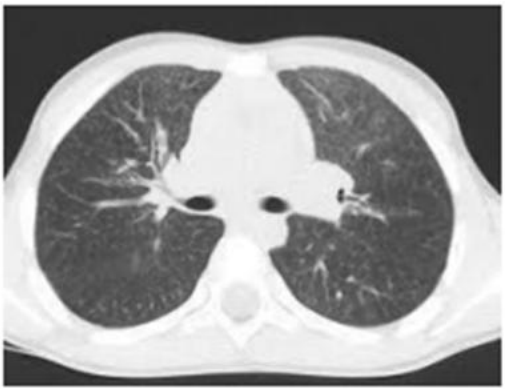

Figure 5: Pre-processing

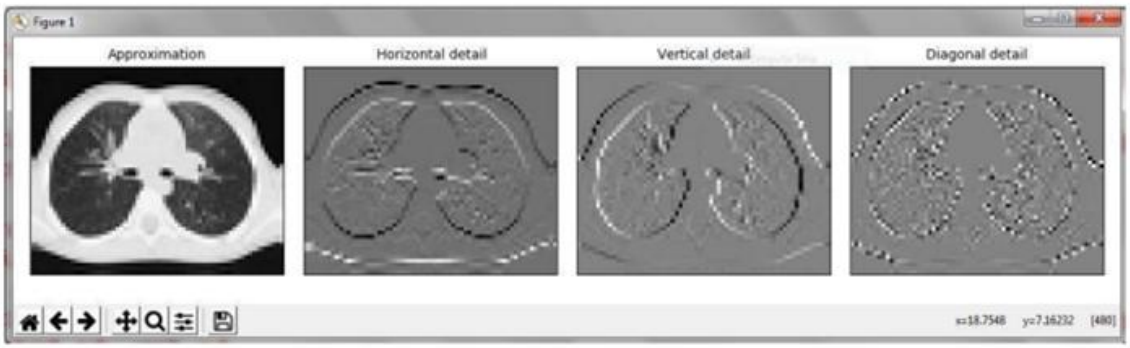

Figure 6: Dual Tree Complex Wavelet Transformation

V. SYSTEM ARCHITECTURE

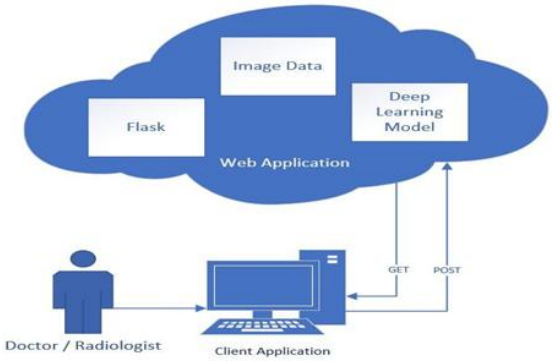

Figure 7: System Architecture

Figure 7 shows a system architecture design. The application is a standard web application that holds the deep learning model and image data that the user uploads via POST request. The user can then user GET request to get images from Flask to either view the CT scan images or view predictions made by the model.

The system utilizes a MVC (Model-View-Controller) software design pattern that splits the code into 3 abstract components: model, view and controller.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 6, pp 222-229, June-2022

https://doi.org/10.47001/IRJIET/2022.606034

5.1 Components

The model in the application is the images that we take from the user and the deep learning model itself. The model drives the main functionality and is central to the entire application. The view in the application is the front-end and is what the user sees, the view uses HTML, CSS, Boot strap and Java Script to display to the user the output of the model.

The controller itself is the Flask back-end code. This is responsible for manipulating data to and from the model and view. In the application, Flask is responsible for the heavy lifting including showing the relevant front-end code through routing to using the model to make predictions.

VI. TESTING

Figure 8: Lung cancer positive

Figure 9: Lung cancer negative

VII. RESULTS AND SNAPSHOTS

7.1 TEST

The procedure starts from collecting the Images on which the model should be trained, then area which should be detected is label for the training set after that the annotation of the images are required which will be followed by pre-processing them. If the number of images are lesser than the required number of images than we can increase the number by image augmentation process. Next, the dataset will be divided into training and testing. Training the model will be done. The model can be ML/DL model but according to the aim DL model will be preferred. The model will be tested in the under testing phase which will be used to detect the lung cancer the uploaded images.

7.2 TRAIN

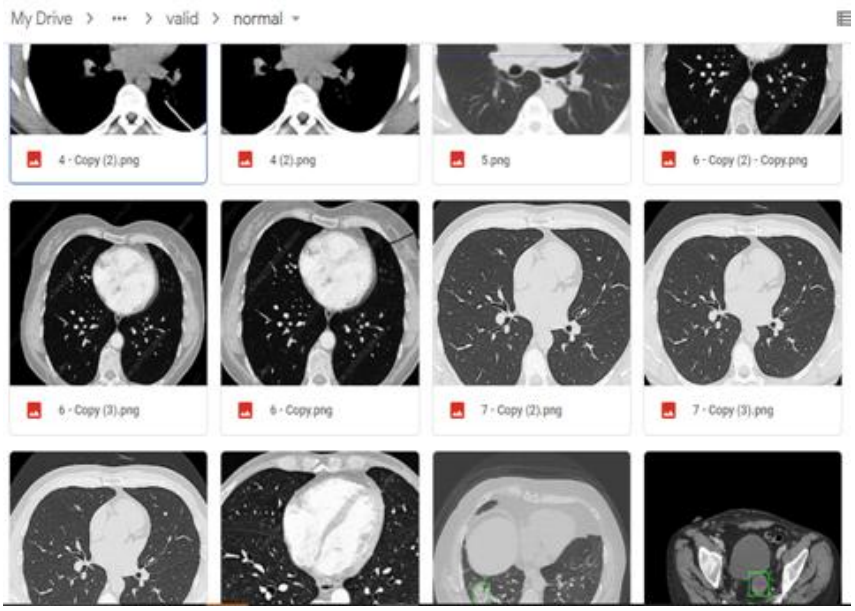

Figure 10: Train Data set

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 6, Issue 6, pp 222-229, June-2022
https://doi.org/10.47001/IRJIET/2022.606034

Figure 11: Test Data set

- Collection of Images in DICOM Format
- Conversion of the images and Labeling the Images
- Annotate all the Images
- Image pre-processing
- Image Augmentation
- Dividing the train and test data set
- Training of the Model
- Testing the Model

7.3 VALID DATA

Figure 12: Valid Data

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 6, pp 222-229, June-2022

https://doi.org/10.47001/IRJIET/2022.606034

Lung cancer is one of the leading causes of death worldwide. There are three major types of lung cancers, non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC) and carcinoid. NSCLC is further classified into lung adenocarcinoma (LADC), squamous cell lung cancer (SQCLC) as well as large cell lung cancer. Many previous studies demonstrated that DNA methylation has emerged as potential lung cancer-specific biomarkers. However, whether there exists a set of DNA methylation markers simultaneously distinguishing such three types of lung cancers remains elusive. In the present study, ROC (Receiving Operating Curve), RFs (Random Forests) and mRMR (Maximum Relevancy and Minimum Redundancy) were proposed to capture the unbiased, informative as well as compact molecular signatures followed by machine learning methods to classify LADC, SQCLC and SCLC. As a result, a panel of 16 DNA methylation markers exhibits an ideal classification power with an accuracy of 86.54%, 84.6% and a recall 84.37%, 85.5% in the leave-one-out cross-validation (LOOCV) and independent data set test experiments, respectively. Besides, comparison results indicate that ensemble-based feature selection methods outperform individual ones when combined with the incremental feature selection (IFS) strategy in terms of the informative and compact property of features. Taken together, results obtained suggest the effectiveness of the ensemble-based feature selection approach and the possible existence of a common panel of DNA methylation markers among such three types of lung cancer tissue, which would facilitate clinical diagnosis and treatment.

VIII. CONCLUSION AND FUTURE ENHANCEMENT

8.1 Model Results

The model resulted in a 65.7% accuracy using the dice coefficient on the training set. The dice coefficient is much lower on the training set however the confusion matrix outputs a high true and false positive rate on a set that contains positive and negative samples. This indicates that the model is great at distinguishing between CT scan slices with no cancer nodules compared to the ones with cancer. I believe with more hyper parameter tuning and model training the accuracy could be increased.

8.2 Concept Evaluation

When doctors find small nodules (less than 3mm) the current practice suggests that they should wait and rescan in 6-12 weeks to see signs of growth. Depending on the tumour, a tumour can grow up to double its size and evolve to a more advanced form of cancer. It is also important to note that the second most frequent diagnosis is small tumours. The project demonstrates that it would be possible for Doctor's to use deep learning applications to aid their decision making process regarding whether a patient with a small tumor should perform a biopsy or rescan in a few weeks which to a patient could mean early treatment and a better prognosis.

8.3 Future Work

Doctors who work in this field are prone to observer fatigue from viewing so many CT scan images. The research on that suggests that observer fatigue increases the risk of errors that can be made by doctors while analyzing these scans. Many images in a CT scan also are irrelevant to Doctors e.g. for 200-300 images only 3 scans would show cancer depending on the stage of the patient. Although this feature was not implemented on the website, a more efficient deep learning model would be capable of alleviating these additional challenges.

8.4 Personal Statement

During the course of the entire project I have learned new skills in areas of deep learning, machine learning, image processing, web development and also research. Being able to blend multiple skills in computer science and produce a proof of concept to try and solve a real world problem is really challenging but also provides the best learning experience. I do believe that anyone who gets involved in Computer Science has a large ability to solve real problems and make the world a better place.

8.6 Justification of project

Anti-spam measures include methods for determining unsolicited email from the email content and methods for using sender information I fit can be determined from the sender's IP address of sender information and the sender's domain name whether the email should be received, it is possible to reduce the processing of the spam filter by the email content that has a high processing load for the determination. This study uses sender authentication technology to identify the sender of forwarded email. We consider that the sender of this forwarded email is the legitimate email sender to receive, and propose to use these as an allow list.

8.7 Future Enhancement

The sender information and authentication mechanism used. Creates an electronic signature from the email header and body using the private key, and adds it to the email as a Signature email header, including related information. Server information described in the record published on server, and authenticates whether it matches the sender of the outgoing mail or not. It uses the IP address of the email source for

ISSN (online): 2581-3048
Volume 6, Issue 6, pp 222-229, June-2022
https://doi.org/10.47001/IRJIET/2022.606034

authentication. Therefore, email from a source different from the original email sender. In this system, sender authentication is performed when receiving mail. From the authentication result, it is used to judge the forwarded mail and extract the sender of the legitimate mail server to build the sender reputation.

REFERENCES

[1] Adam: A Method for Stochastic Optimization, URL https://arxiv.org/abs/1412.6980v8
[2] Analyzing The Papers Behind Facebook's Computer Vision Approach Adit Deshpande CS Undergraduate UCLA ('19),. URL https://adeshpande3.github.io/Analyzing-the-Papers-Behind-Facebook%27s-Computer-Vision-Approach/
[3] Computer Vision - deep learning. ai, URL https://www.coursera.org/learn/convolutional-neural-networks/lecture/Ob1nR/computer-vision
[4] Data Science Bowl 2017 Kaggle, URL, https://www.kaggle.com/c/data-science-bowl-2017/kernels
[5] Floyd Hub- Deep Learning Platform- Cloud GPU, URL https://www.floydhub.com
[6] Houns field unit,. URL https://medical-dictionary.thefreedictionary.com/Hounsfield+unit
[7] Hounsfield Scale LITFL Life in the Fast Lane Medical Blog, URL https://lifeinthefastlane.com/funtabulously-frivolous-friday-five-164/screen-shot-2016-10-14-at-11-19-30/
[8] LUNA16 Grand Challenge, URL https://luna16.grand-challenge.org/
[9] Lung cancer: Reduce your risk by quitting smoking, URL http://www.mayoclinic.org/diseases-conditions/lung-cancer/basics/definition/con-20025531
[10] Manifestofor Agile Software Development, URL http://agilemanifesto.org/
[11] Neural Network Foundations, Explained: Activation Function, URL https://www.kdnuggets.com/2017/09/neural-network-foundations-explained-activation-function.html
[12] Object Localization and Detection Artificial Intelligence, URL https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/object_localization_and_detection.html
[13] Overview of neuron structure and function, URL https://www.khanacademy.org/science/biology/human-biology/neuron-nervous-system/a/overview-of-neuron-structure-and-function
[14] Project Jupyter, URL http://www.jupyter.org
[15] Srensen-Dice similarity coefficient for image segmentation- MATLAB dice-Math Works United Kingdom, URL https://uk.mathworks.com/help/images/ref/dice.html
[16] Tensor Flow, URL https://www.tensorflow.org/
[17] Welcome to Python.org, URL https://www.python.org/
[18] Welcome Flask (A Python Micro frame work), URL http://flask.pocoo.org/

AUTHORS BIOGRAPHY

Anusha S, UG Student, Department of Computer Science & Engineering, Vidya Vikas Institute of Engineering and Technology, Mysore, India.

Chandrakumar, UG Student, Department of Computer Science & Engineering, Vidya Vikas Institute of Engineering and Technology, Mysore, India.

Gunashree K, UG Student, Department of Computer Science & Engineering, Vidya Vikas Institute of Engineering and Technology, Mysore, India.

Suma N, UG Student, Department of Computer Science & Engineering, Vidya Vikas Institute of Engineering and Technology, Mysore, India.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 6, pp 222-229, June-2022

https://doi.org/10.47001/IRJIET/2022.606034

Citation of this Article:

Anusha S, Chandrakumar, Gunashree K, Suma N, Dr. Madhu B K, "Lung Cancer Detection Using Machine Learning" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 6, Issue 6, pp 222-229, June 2022. Article DOI https://doi.org/10.47001/IRJIET/2022.606034

****

