IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

# Implementation of Lung Cancer Detection & Recommendation of Oncologist Using Machine Learning

**Nachiket Kelkar [1*], Niraj Mate [2], Atharv Kukade [3], Abhijit Kulkarni [4], Pradnya Mehta [5]**

Department of Computer Engineering, MMCOE, Karvenagar, Pune, India [1, 2, 3, 4]

Department of Computer Engineering, MMCOE, Karvenagar, Pune, India[5]

[*]*Corresponding Author: kelkarnachiket05@gmail.com, Tel.: +91-7588276423*

DOI: https://doi.org/10.26438/ijcse/v7i5.467471 | Available online at: www.ijcseonline.org

Accepted: 12/May/2019, Published: 31/May/2019

***Abstract*—** Lung cancer is one of the most prominent and deleterious forms of cancer and affects about 2lakh people every year on an average. On a positive note, Lung Cancer death rates have significantly declined over the past decade due to early detection and treatment. Hence, this system uses CT images for detection of lung cancer. It contains several steps like image acquisition, pre-processing, thresholding, segmentation, feature extraction and detection of the presence and the stage of cancer if it is present. Initially, unwanted noise is removed using filters. In the next step, thresholding is used to perform segmentation and highlight the tumour spots. Using flood fill, and masks on the thresholded image, tumour spots which are isolated from the rest of the image are obtained. Features like area, perimeter and number of tumour spots, etc. are extracted by calculating contours using edge detection. Extracted features are given to the classifier model to detect the presence and hence the stage of existing cancer. The system then goes ahead and generates a report which is sent to the doctor for further analysis.

***Keywords*—** Image Processing, Machine Learning, Preprocessing, Binarization, Segmentation, and Feature extraction.

## I. INTRODUCTION

With the rising trends in technology, almost everything is being done over the internet. Many healthcare related web applications and websites are coming forward like TreatHF, Preconception care, Mayo Clinic etc. This benefits many individuals seeking preliminary medical assistance at the tip of their fingers before receiving proper help from the appropriate medical practitioner. Data analytics and Machine Learning can also assist doctors by analysing patient data and giving a prediction based on it. That being said, this rise in technology in no way indicates that it can replace the proper trained professionals in any way. All of these technologies can be seen more as tools to make the process of medical diagnosis and healthcare easier for the patients as well as the concerned physicians. This is a web-based system for the detection of lung cancer using the CT scan image that is uploaded by the patient onto the site. On the server end this image goes through various steps and a report is generated based on it which is sent to the doctor. The purpose of this report is to serve as a reference to the doctor to aid with the diagnosis procedure and provide proper consultation to the concerned patient.

## II. RELATED WORK

In the papers referred for this survey, Convolutional Neural Network (CNN) based on deep learning technique is used for training patients' dataset. CNN having the advantage of both classification and prediction of input serves a robust and reliable algorithm to train against large datasets as CNN has N number of parameters to be manipulated for accurate training. However, this paper only classifies images into whether the cancer is present or absent [1]. Using neural networks for classification the design of the system was observed to be ideal for performing as a Computer Aided Detection (CADe) system for detection of lung cancer [2]. An optimized Lung cancer detection system approach shows that, Lung Cancer Detection from CT images can be done in four stages. In this case BAT Algorithms are used to improve performance system and ANN is applied to revamp accuracy of results [3]. For reference on using image processing techniques this paper was referred which presents in detail literature survey on various techniques that have been used in Pre-processing, nodule segmentation and classification [4]. In one of the proposed approaches, image preprocessing techniques used like median filter for noise removal, High boost operator for enhancement and marker-controlled

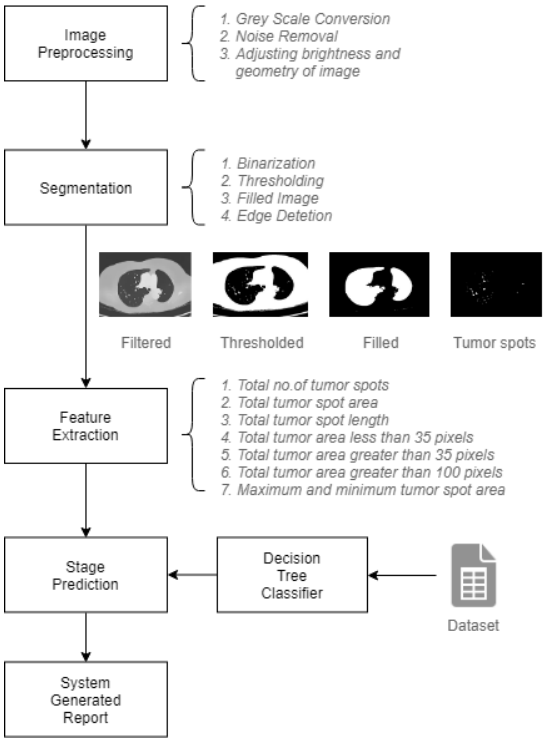
 

watershed used for image segmentation then the feature extraction give suspicious region of interest of tumor. Finally describes the various classification techniques for detecting lung cancer [5]. Using the genetic approach, it can be observed that manual interpretations are time consuming and very critical, to overcome this difficulty the best features of Genetic Algorithm method and Naive Bayes Classification are taken to classify the different stages of the cancer images fast and accurately [6]. One of the proposed systems in a research paper that was referred is designed to detect lung cancer in premature stage in two stages. In first stage, Binarization technique is used to convert binary image. In second stage, segmentation is performed to segment the lung CT image. The proposed system is designed such that it can detect which lung is affected left lung or right lung specifically [7]. The different stages in a Computer Aided Lung Cancer Detection system are Enhancement, Segmentation and Feature extraction. There are different techniques for performing these stages. These different techniques have been explained in detail [8]. In a few of the papers referred, MATLAB was used to perform image processing tasks on the CT image and this research paper provides an alternative approach to it. In image processing steps, processes such as image pre-processing, segmentation and feature extraction have been discussed in detail [9].

## III. METHODOLOGY

The purpose of the proposed system is to aid the physicians/oncologists in the procedure for diagnosis of lung cancer. A large dataset containing various images for cancer research purposes is used to train the machine learning model in this case a decision tree classifier in order to identify the stage of cancer if present. When an image is uploaded by the patient onto the system it undergoes various preprocessing steps in order to remove unnecessary noise, a mean shifted filter is used to remove the noise present in the CT image. Then the image is converted to grey scale the image is then thresholded, there are basically two methods for thresholding i.e. otsu thresholding or manual thresholding, in this system manual thresholding is used, further by using flood fill algorithm in combination with masking technique the tumor cells on the CT image are detected, then contours are used in order to extract the required features of cancerous cells/tumor spots. This is done using OpenCV (Open Source Computer Vision Library) which contains interfaces to support image processing operations. Various features are extracted from the image like the number of tumor spots, total area of the spots, area of the biggest spot, area of smallest spot, total number of tumor spots, area of tumor spot greater than thirty, area of tumor spots greater than hundred etc.

**The system architecture is as shown below:**




## IV. RESULTS AND DISCUSSION

### a. Preprocessing

First step after taking the CT image as input is to apply pre-processing on it. The input CT image may contain errors related to the brightness values, geometry, and lack of contrast of the pixels. Following preprocessing steps applied to input CT image to convert the image to a form which is better suited for machine interpretation.

1. Grey Scale Image

**Fig 1. Original Image**

It is also called as color filtering which means to convert RGB image frames into HSV (Hue Saturation Value) image. The image is further converted in a gray scale image. Gray scale images are logically a matrix with each pixel represented as a discrete level out of 0 to 255 [10]. Gray Scale also helps to remove hardware errors if any.

2. Noise Removal

Noise Filtering will remove the unnecessary information from an image. It is also used to remove various types of

noises from the images. Noise removal uses filters like low pass, high pass, mean, median etc.
In this system a mean shifted filter is used to remove the unwanted noise from the CT image. Refer Fig1 & Fig2.

**Fig 2. Grey scaled Image**

**Fig 3. Mean Shifted Image**

### b. Segmentation

1. Binarization

After preprocessing the CT image the image is thresholded, there are basically two methods for thresholding i.e. otsu thresholding or manual thresholding, in this system manual thresholding is used Binarization will convert the gray scale CT image to a binary image. The binary image contains only two pixels black and white.

**Fig 4. Thresholded Image**

Here the tumors in the CT image lie on the higher side of the spectrum so the threshold value is set based on that. The pixels with tumor spots and other places having greater value than threshold will be converted to the value 1, while the others will be neglected and converted to 0.

2. Dilation and Filled Image

Dilation is used to add pixels at region of boundaries or to fill in holes which generate during erosion process. Dilation can also be used to connect disjoint pixels and add pixels at edges.

**Fig 5. Filled Image**

The dilated image is further converted into a filled image which highlights the areas which are important. A floodfill algorithm with masking is used to find the tumor spots present in the CT image. A thresholded image is masked by a filled lung image.

3. Edge Detection/Tumor spots detection

After the masking is done the output of the stage is the tumor spots present on the lungs. Contours are then used to find various features of the tumor spots which are extracted from the previous step.

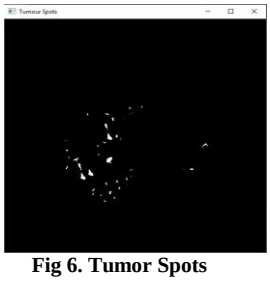


**Fig 6. Tumor Spots**

The contours are a useful tool for shape analysis and object detection and recognition. Contours basically use edge detection technique to find the spots, for the contours to work better the image should be in binary format, as here the image is converted in the binary format already contours work efficiently. Generally edge detection methods transform original images into edge images. The edge representation of an image significantly reduces the quantity of data to be processed, yet it retains essential information regarding the shapes of tumor spots in the binarized CT image.

### c. Feature Extraction

After the tumor spots are obtained after processing the image the next step is to extract the features and predict the stage. This is a crucial and essential stage after image segmentation. The normality or abnormality of an image is determined by the final results of feature extraction stage.
Here the features are extracted using the contours in opencv. The contours are a useful tool for shape analysis and object detection and recognition. For better use of contours the image should be in binary format. The final output of the image processing stage is in binary format and hence contours can be used. The contours has many properties which are helpful for finding the area, perimeter and other features of the tumor spots.
Extracted features works as the base for classification process. Following features will be extracted:

1. Number of tumor spots.
2. Total area of the spots.
3. Area of the largest spot.
4. Height of the largest spot.
5. No of tumor spots having area less than 30

6. No of tumor spots having area greater than 30
7. No of tumor spots having area greater than 100

The features are defined as follows:

1. Number of tumor spots: It is a numerical value which is taken after calculating total number of tumor spots in the binarized CT image.
2. Total area of the spots: This value is retrieved by the summation of areas in the image which are marks as 1 in the binary CT image.
3. Area of the largest spot: This is the numerical value which is obtained after identifying the largest spot in the CT image.
4. Height of the largest spot: After the spot identification which is largest of all the spots in the CT image, height is calculated.
5. No of tumor spots having area less than 30 : The count of tumor spots in the processed having the area less than 30 pixels.
6. No of tumor spots having area greater than 30 : The count of tumor spots in the processed having the area greater than 30 pixels
7. No of tumor spots having area greater than 100 : The count of tumor spots in the processed having the area less than 100 pixels.

### d. Prediction

After the feature extraction we get the features of the of the tumor spots present in the CT image which helps us to do the prediction of stage of cancer. A decision tree classifier is used in this system for prediction stage. A well-defined dataset with attributes/features with stages is used to construct the decision tree. Further this decision tree is used for prediction purposes.

A Decision tree classifier is a widely used classification technique. It is simple to implement. It applies straightforward idea to solve a classification problem. It tries to solve the problem by using tree representation. Each internal node of the tree represents the attribute while the leaf node corresponds to the class label. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached.

Here after the tree was constructed the TotalAreaOfTumorSpot attribute was considered as the root node of the tree. The accuracy of the decision tree classifier is found out to be >85 percent.

## V. ANALYSIS

The proposed system was tested on the available dataset of 104 DICOM images having cancer (with stage I to IV) and no-cancer.

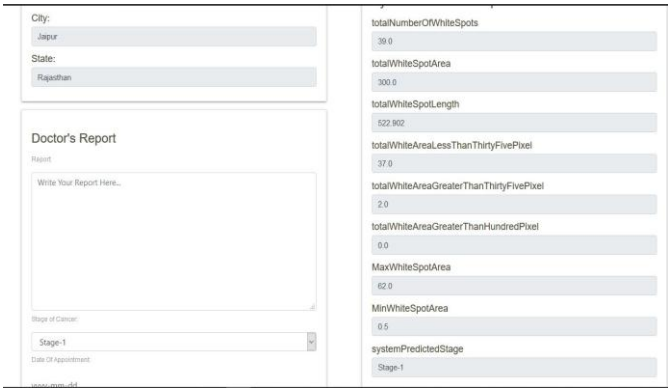



**Fig7. System Output/Doctor's Report**

The above image is the generated report of the patient. It contains the patient details, the extracted features of the CT image uploaded by the patient, the ct image and the stage predicted by the system. The performance evaluation of the proposed method resulted in detection accuracy of 83.69%. The accuracy can be increased if the size of the dataset to train the classifier is increased. The confusion Matrix is given below:

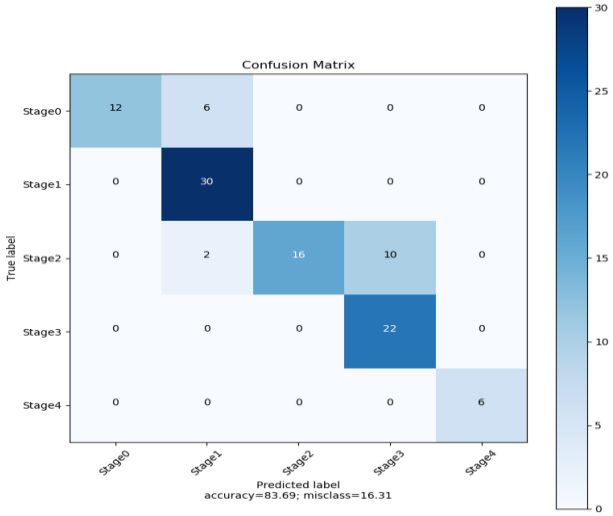



**Fig8. Confusion Matix**

The confusion matrix represents the result of classification, The diagonal elements represent the correctly classified images while the other cells represent the misclassified images. 84 images were classified correctly while the others were misclassified.

Accuracy of the model is calculated as:

**Accuracy = (TP + TN) / (TP + TN + FP + FN)**

Accuracy = correctly classified elements/ total elements

The TN and TP represent the elements classified correctly.

The accuracy of this model is 83.69%.

A decision tree classifier was used in the system for classification of the new data entries. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached.

The decision tree constructed after training the dataset is as given below:

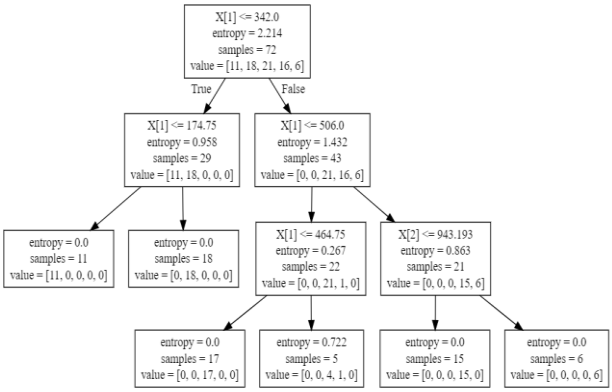



**Fig9. Decision Tree**

The basic idea of the decision tree is that it tries to solve the problem by using tree representation. Each internal node of the tree represents the attribute while the leaf node corresponds to the class label. The attribute with the highest entropy is placed on top. Here the attribute which is considered superior for classification is the total white spot area attribute. X represents the attribute, entropy represents the entropy at that internal node, and samples represent the number of samples present at that stage or node of the tree.

## VI. CONCLUSION AND FUTURE SCOPE

The proposed system will be able to detect not only the presence of lung cancer in a patient's CT scan image but also classify it according to its stage and generate a detailed report for the concerned physician to see. A dataset of many cancer patients with varying levels of the disease is used to train the learning model. It aims to provide a web-based platform for patients to upload their image and to get their report generated and sent to their doctor to get further analysis.

In the future scope of this project, we can expand on the current system to track the patient's consultations with the doctor and see the yielded results. In this way the system can become more involved with the patient by fetching more data on the patient through various other sources.

## REFERENCES

[1] Nachiket Kelkar**,** Niraj Mate**,** Abhijit Kulkarni**,** Atharv Kukade**,** Pradnya Mehta**,** "Lung Cancer Detection & Recommendation of Oncologist using Machine Learning", JETIR 2018.

[2] Karan Sharma, Harshil Soni and Kushika Agarwal, "Lung Cancer Detection in CT Scans of Patients Using Image Processing and Machine Learning Technique", Springer 2018.

[3] S.Kalaivani, Pramit Chatterjee, Shikhar Juyal, Rishi Gupta, "Lung Cancer Detection Using Digital Image Processing and Artificial Neural Networks.", ICECA 2017

[4] Sheenam Rattan, Sumandeep Kaur, Nishu Kansal, Jaspreet Kaur, "An optimized Lung Cancer Classification System for Computed Tomography Images.", IEEE 2017.

[5] G.Niranjana, Dr.M.Ponnavaikko, "A Review on Image Processing Methods in Detecting Lung Cancer using CT Images.", ICTACC, 2017.

[6] Pooja R. Katre, Dr. Anuradha Thakare, "Detection of Lung Cancer Stages using Image Processing and Data Classification Techniques." I2CT 2017.

[7] Mansee Kurkure, Anuradha Tharkre, "Lung Cancer Detection using Genetic Approach." 2017

[8] Md. Badrul Alam Miah, Mohammad Abu Yousuf, "Detection of Lung Cancer from CT Image Using Image Processing and Neural Network," ICEEICT, 2015.

[9] Sruthi Ignatious, Robin Joseph, "Computer Aided Lung Cancer Detection System." GCCT 2015.

[10] Anita Chaudhary, Sonit Sukhraj Singh, "Lung Cancer Detection On CT images by using image processing" ICCS, 2012.

[11] Gayatri. D. Patil, Lubdha. M. Bendale, Roshani. L. Jain, "*Document Image Noises and Removal Methods*", International Journal of Scientific Research in Computer Science and Engineering, Vol.06, Issue.01, pp.48-63, 2018

[12] Roshani. L.Jain, Lubdha M. Bendale, Gayatri D. Patil, "*Image Enhancement Using Different Techniques*", International Journal of Scientific Research in Computer Science and Engineering, Vol.06, Issue.01, pp.73-76, 2018

## Authors Profile

*Mr. Nachiket Kelkar*, pursued Bachelor of Engineering in Computer Science from Savitribai Phule Pune University, India in 2019. He is a member of ACM since 2015. He is currently employed at Persistent Systems as a Software Engineer. His areas of interest include Machine Learning, Data Science and Data Analytics.

*Mr. Abhijit Kulkarni*, pursued Bachelor of Engineering in Computer Science from Savitribai Phule Pune University, India in 2019. He is a member of ACM since 2015. He is currently employed at Larsen & Toubro Infotech as a graduate engineering trainee. His areas of Interest Include Cyber Security, Cloud Security and Privacy and Big Data Analytics.

*Mr. Atharv Kukade*, pursued Bachelor of Engineering in Computer Science from Savitribai Phule Pune University, India in 2019. He is a member of ACM since 2015. He is currently employed at Ellicium as a Senior Software Developer. His areas of Interest include Machine Learning, Data Science and Data Mining.

*Mr. Niraj Mate*, pursued Bachelor of Engineering in Computer Science from Savitribai Phule Pune University, India in 2019. He is a member of ACM since 2016. He is currently employed at Infosys ltd. as a Software Engineer. His areas of Interest include Data and Cyber Security, IOT and Big Data Analytics.

*Ms. Pradnya Mehta*, currently working as an Assistant Professor at Marathwada Mitra Mandal's College of Engineering, Pune.