Detection of Lung Cancer using Image Processing Techniques

Madhusudhana Rao Dontha[1], Praneeth Cheraku[2], G.Appa Rao[3] SubbaRao Peram[4]

[1,4] Department of IT, Vignan's Foundation for Science, Technology & Research, Vadlamudi, AP, India.

[2] Department of IT, Prasad V.Potluri Siddhartha Institute of Technology, Vijayawada, AP,India

[3] Department of CSE, Avanthi's St.Theressa Institute of Engineering & Technology, Garividi, AP,India

madhu.dontha@gmail.com,chpraneeth@hotmail.com, subbarao.peram@gmail.com

Abstract: Image processing techniques are used mostly in many medical areas, where the time factor is very important to discover the abnormality in the target image. As we all are aware, lung cancer is one of the most dangerous and life-taking diseases in the world, recognizing it in the early stages is vital. Though CT scan imaging is the best imaging technique in the medical field, the doctors find it difficult in identifying the lung tumor in the CT scan images due to which the computer-aided diagnosis came into existence. As a result, computer-aided diagnosis (CAD) makes it simple for doctors to precisely identify malignant cells. Many image processing and machine learning-based computer-aided approaches have been developed and applied. We proposed a method to detect cancer tumours in the lungs based on the area in this study. This paper mainly deals with the identification of the lung cancer stage. Initially, the acquired CT scan image is preprocessed and segmented. In the later stages, we compute the area of the separated ROIs and classify them based upon it. The result shows the tumor region and to which type does it belong, whether it is cancerous or non-cancerous.

***Keywords**:* image processing techniques; CT scan imaging; Computer-Aided Diagnosis (CAD); image preprocessing and segmentation, ROI.

I INTRODUCTION: The formation of abnormal cells lining the air passageways inside the lung tissue is referred to as lung cancer. These cells proliferate and develop at a faster rate than normal cells, forming a cluster or tumour. These days lung cancer has become one of the major causes of death. Lung cancer has the highest mortality rate of all cancers in this world. Lung cancer survival is affected by the tumour size at the time of diagnosis. Non-small cell lung cancer and small cell lung cancer are the two types of lung cancer. These lung cancer kinds are classified based on their characteristics. Lung cancers are classified into four types: I, II, III, and IV. The size of the tumour can be used to stage it.

According to World Health Organization (WHO) data, cancer is the biggest cause of mortality, accounting for an estimated 9.6 million deaths in 2018. Lung cancer is at the top of the list, with an estimated 2.09 million cases and 1.76 million deaths. Lung cancer is responsible for 21% of all cancer-related deaths. In the year 2019, 228,150 new cases are expected, with 142,670 expected to die, accounting for 62 percent of the death rate. Furthermore, lung cancer death rates are predicted to rise, with 21 million deaths expected by 2032.

It’s very important to keep in mind that most of the cancer-affected people in developing countries are poor. So, if they want to get diagnosed, they need to pay hefty consultation charges. In light of this, we created a computer-assisted diagnosis system that makes cancer detection more economical.

In recent years, there have been many reported researches on lung cancer detection. In 2011, Disha Sharma and Gagandeep Jindal [1] built a CAD system which involve some basic image processing techniques such as Erosion, Median Filter, Dilation, Outlining and Lung Border Extraction. In [2] proposed a system which uses convolution filters for image enhancement and thresholding for segmentation. Later computed features like area, perimeter and roundness using morphologic and colorimetric techniques. A.Amutha and Dr.R.S.D.Wahidabhanu [3] proposed an approach for detecting and segmenting lung cancer. They began by acquiring the lung CT picture and preprocessing it with the denoising algorithm. Secondly, the extracted noise, free image is given for the feature extraction process and later segmented using level set-active contour model. Prof. Anuradha S. Deshpande [4] proposed imaging processing technique for lung cancer detection. In this paper, the CAD system uses SVM classifier in order to identify the cancer and to classify it in the different stages of cancer. Firstly, the image is acquired which is a fusion of CT scan and MRI images. Noise reduction and image enhancement techniques are applied on it. The preprocessed image is segmented using watershed technique. Features like area, perimeter and eccentricity are extracted from the segmented image and at last SVM classifier is used for classification. SnehalDabade et al. [5] created a CAD framework for capturing and enhancing lung CT scan images with FFT, Auto enhancement, and Gabor filtering The image is then segmented using a thresholding technique. On the segmented image, the watershed approach is used. The image's area, perimeter, average intensity, and roundness are then calculated. Finally, based on these characteristics, the malignancy was detected. Ayushi Shukla [6] and team suggested a detection method that uses multiple filtering techniques such as averaging filters, median filtering, and morphological filtering to preprocess the acquired lung CT scan picture. Segmenting the preprocessed image entails procedures such as thresholding, edge detection, and watershed transform. The segmented picture is used to extract features such as perimeter, area, and eccentricity, which are then categorised using SVM. JaneeAlam [7] presented a CAD algorithm that uses the masking approach to acquire and enhance an image. The features are extracted from the preprocessed image using the watershed technique. Mean, standard deviation, entropy, RMS, fluctuation, smoothness, kurtosis, IDM, differentiation, relationship, vitality, and homogeneity are some of the elements used for extraction. SVM classifier is employed for classification, and total lung area and total disease affected area are computed. The fundamental drawback of these methods is that detection is limited to segmentation. These approaches also have some flaws that can be addressed by upgrading the technology utilised in them. Furthermore, all the work has done only for detecting whether it’s cancerous or not but none detected to which stage does the cancer belong. The main contributions of our work are summarized as follows:

We attempted to overcome these issues with our proposed approach. Our suggested method can identify the tumour as well as the cancer stage (I, II, III, or IV). If no tumours are discovered in the provided image, the result is No Cancer.

The remainder of this work is organised in the same way that Section II discussed the suggested algorithm's technique. We examined the experimental data in Section III. Section IV discusses the conclusion and future work.

II. METHODOLOGY

A coherent computer aided diagnosis system is built for lung cancer detection using MATLAB, where we use image processing technique. The proposed method for the lung cancer detection follows three basic diagnostic tasks of radiology namely, preprocessing, segmentation and feature extraction. We compute and classify the area based on it in a later stage. As previously indicated, the CT scan image is preprocessed. The preprocessed image is segmented, and the characteristics from the segmented image are extracted subsequently. Finally, we use the retrieved features to classify the image. The overall flowchart of the proposed system is depicted in the fig.1.

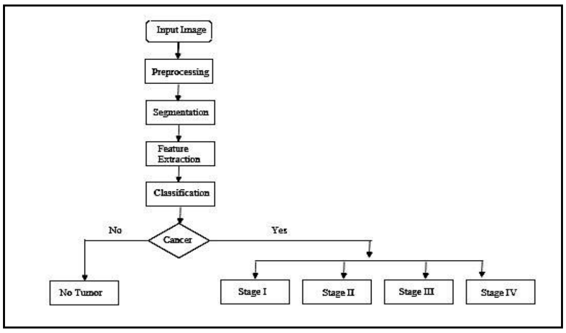

Fig.1. Flowchart of the proposedsystem

i. Image Preprocessing

At this point, we're mostly concerned in image improvement. The primary goal of image enhancement is to improve the impression of the image's data. As a result, the results are better suited for additional picture processing. As shown in fig.2, we enhanced the image in four steps in our suggested method.

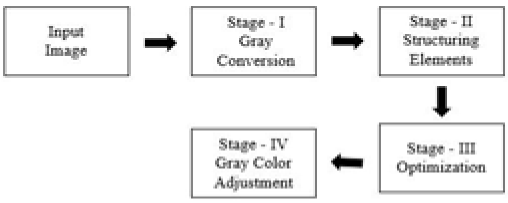

Fig.2. Flowchart showing the preprocessing techniques

Now, let us see the order the things happen in this stage i.e., Image Preprocessing.

- *Stage – I*

In this stage, the acquired input image is converted into grayscale to get the better result. We have used *rgb2gray* function in MATLAB.

- *Stage – II*

In this stage, the elements in the gray scale image acquired from the stage-I are morphologically structured. It is majorly used in morphological dilation and erosion operation. We have used *strel* function in MATLAB.

- *Stage – III*

In this stage, optimization is given the paramount importance. Here, we have subtracted the image obtained in the stage-II from the image acquired in the stage-I.

- *Stage – IV*

In this stage, we have fine-tuned the intensity values in the grayscale image. This operation was performed to increase the contrast value. We have used *imadjust* function in MATLAB.

ii. Image Segmentation

This stage plays imperative role in the image processing techniques. As the word segmentation itself tells the image is divided into different segments. Here, the division into parts is done based upon the characteristics of the pixels so that we can find the tumor region. Thus, the preprocessed image is converted into a binary image to get the tumor part.

iii. Feature Extraction

The next stage entails determining the properties of the items in the image. This is called as feature extraction. In image processing, feature extraction is noteworthy. Feature extraction delineates the pertinent shape information contained in the image so that it becomes easy at the time of classification. It is useful when we want to reduce the redundant information from the image.

We used *graycoprops* function in the MATLAB. We computed four features contrast, correlation, energy and homogeneity. As these features are in the neighbourhood of each other so it became strenuous. Keeping this in mind we computed another feature i.e. area. We constructed correlation graph of each and every feature with the other. And we came to know that area has the strongest relationship with the cancer stage as shown below in the fig.4. So, we will classify an image based upon the area of the tumor. Area can also be helpful for prediction of the cancer as we have to know the rate at which the tumor grows. We have also lay out a heat map so that envision the areas of location as manifested in the fig.3.

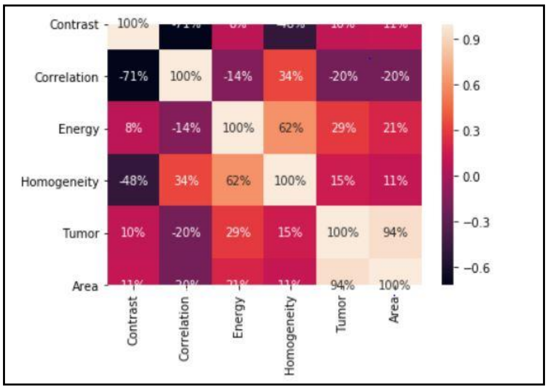

Fig.3. Heat map of evaluated features

Now, let us know the significance of the features which we have computed.

- *Contrast*

Inertia is another name for the property contrast. Over the entire image, it delivers the intensity contrast between a pixel and its neighbour. The formula for calculating contrast is

Contrast =

- *Correlation*

This feature returns the correlation between a pixel and its neighbours over the entire image. The correlation was calculated using the following formula:

Correlation = $\sum_{i,j=0}^{N-1} \rho_{ij} ((i-\mu)(j-\mu)),$

- *Energy*

Uniformity is another name for the property energy. In the GLCM, it returns the sum of squared elements. The formula used to calculate the energy is as follows:

Energy =

- *Homogeneity*

This property returns a value that indicates how near the GLCM's element distribution is to the GLCM diagonal. The formula for calculating homogeneity is

Homogeneity = $\sum_{i,j=0}^{N-1} (\rho_{ij})/($

- *Area*

This feature plays a salient role in this CAD system. As the above computed features are closely related to each other so we took area into consideration. We have calculated the number of white pixels of the binary image by using the *nnz* function which will return the number of non-zero elements.

iv. Cancer Stage Classification

By scrutinizing the features which we have extracted from the lung CT scan image, we have found that the features computed by using the GLCM cling so closely to each other as shown below in the fig.3. That’s the reason we took area into consideration and the classified the image. Here, we have 5 target classes and they are 0, 1, 2, 3, and 4; where class 0 represents no tumor, class 1 represents cancer stage-1, class 2 represents cancer stage-2, class 3 represents cancer stage-3 and class 4 represents cancer stage-4.

After quantifying the area of the tumor, we had classified the obtained image. We have examined all the 130 lung CT scan images and their respective area. Later we set a range for each and every stage.

if the area >= 7700 and area < 10000 then its CANCER STAGE-1

else if the area >= 10000 and area <= 15000 then its CANCER STAGE-2

else if the area >= 15000 and area <= 25000 then its CANCER STAGE-3

else if the area >= 25000 then its CANCER STAGE-4

else there is NO CANCER

III. EXPERIMENTAL RESULTS AND ANALYSIS

The lung picture set exploited for the probing of the proposed algorithm comprehend 130 lung CT JPEG pictures in which 50 are non-cancerous and 80 are cancerous, which is again classified into 4 stages i.e., Cancer stage-I, cancer stage-II, cancer stage-III and cancer stage-IV. We have 20 lung CT JPEG picture for each stage of cancer. In output, we have displayed 3 images. The acquired input image is the first image. The second image is a grayscale image that was created using our proposed system and all of the stages of the preprocessing technique. Finally, there's the binary image, which was created using the segmentation process.

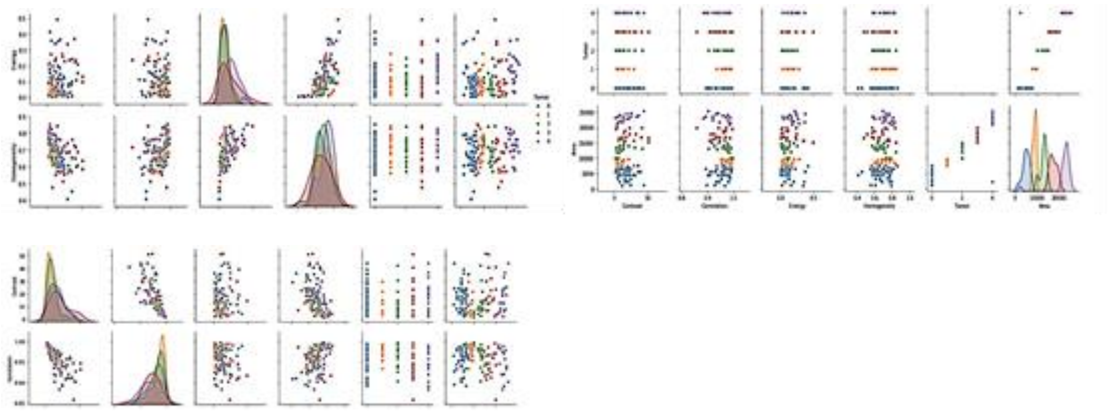

Fig: 4 Computation of Contrast, Correlation, Energy and Homogeneity, Area using MATLAB

This image divulges the tumor in the acquired input image with a help graphs for each of the statistical feature i.e., Contrast, Correlation, Energy and Homogeneity; which are extracted for the tumor in the segmented image. If the acquired input image is cancerous then the graphs contain a spline and if the acquired input image is non-cancerous then the graphs will be empty as there is no tumor.

The findings obtained for a normal lung, i.e., a lung with no tumour, are shown in fig.5, and the graph for it is given in fig.6. The results obtained for an abnormal lung, i.e., a lung with a tumour that leads to cancer, are shown in fig.7, and the graph for it is shown in fig.8.

Our test dataset contains 50 lung CT JPEG images. There are 10 images each of No cancer stage, Cancer stage-1, Cancer stage-2, Cancer stage-3 and Cancer stage-4. Out of 40 cancer-affected photos, the suggested algorithm correctly identified all 40 as cancerous and 0 as non-cancerous. It also correctly identified the cancer stage with a 100% accuracy rate. When it comes to non-cancerous photos, which number ten, the suggested algorithm was able to correctly identify all ten as non-cancerous and 0 as cancerous, resulting in a 100% accuracy. This is represented in the Table I. We also performed some operations on our dataset in the Jupyter Notebook in python to generate some reports. In jupyter notebook, we bifurcated our dataset into training dataset with 80 images and testing dataset with 50 images as we have total of 130 images in our dataset. Table II lists the accuracy of three well-known machine learning algorithms: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier, as well as their training and testing datasets. Table III shows a comparison between previous work and our suggested approach. When compared to previous systems, the accuracy obtained by this approach is higher.

Fig.5. Output result for a non-cancerous lung image

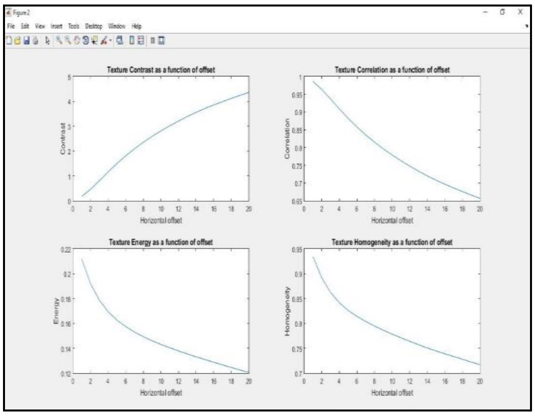

Fig.6. Graph plotted for the statistical features for the image acquired in fig.5.

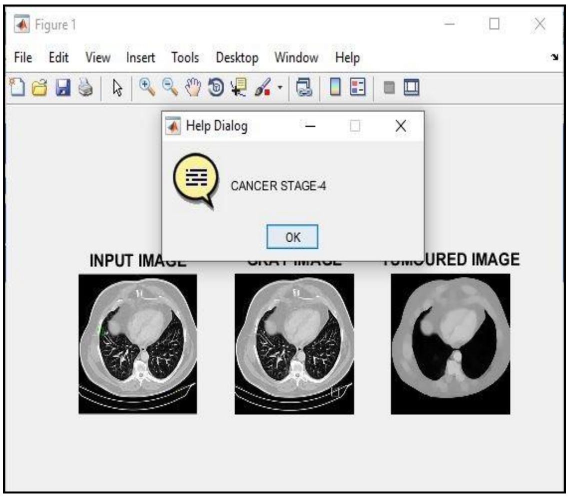

Fig.7. Output result for a cancerous lung image of stage-4

TABLE I ACCURACY FOR DETECTION

IMAGE TYPE	Description			
	No. of Images	Detected Images	Undetected Images	Accuracy in %
Non-Cancerous	10	9	1	90
Cancerous	40	37	3	92.5

TABLE II ACCURACY OF DIFFERENT ALGORITHMS

DATASET	Accuracy in %		
	Logistic Regression	Decision Tree Classifier	Random Forest Classifier
Training	92	100	98
Testing	89	100	95

TABLE III COLLATION BETWEEN EXISTING WORKS AND OUR PROPOSED ALGORITHMS

AUTHOR	YEAR	TECHNIQUES APPLIED	ACCURACY IN %
K. Murphy et al.	2009	Region Growing and Morphological Smoothing	84
Disha Sharma and Gagandeep Jindal	2011	Erosion, Median Filter, Dilation, Outlining, Lung Border Extraction	80
S. K. Kumar et al.	2011	Bio-orthogonal Transform and Region Growing	86
Prof. Anuradha S. Deshpande and her students	2015	Noise Reduction and Watershed Algorithm	85
Ayushi Shukla et al.	2018	Filtering, Thresholding, Edge Detection & Watershed Transform	82
Our Proposed Algorithm	**2020**	**Grayscale Conversion and Binary Image Conversion**	**92**

IV. CONCLUSION AND FUTURE WORKS

In this research paper, we offer an algorithm that can identify lung cancer and provides unco results, in contrast to other commonly used methods. From the isolated region of interest, a set of features is unsheathed (ROI). This method may readily detect a tumour and its associated cancer stage, such as I, II, III, or IV. If no tumours are discovered in the provided image, the result is No Cancer. It has a 92 percent accuracy rate. The proposed technology would be useful in assisting doctors in determining if an image is malignant or not. In the future, we will extemporise by considering all of the extracted attributes and then classifying based on them.

REFERENCES

[1] Disha Sharma, Gagandeep Jindal, "Identifying Lung Cancer Using Image Processing Techniques", International Conference on Computational Techniques and Artificial Intelligence, pp. 116-120, 2011.

[2] A. Chaudhary and S. S. Singh, "Lung Cancer Detection on CT Images by Using Image Processing," *2012 International Conference on Computing Sciences*, Phagwara, 2012, pp. 142-146.

[3] Amutha and R. S. D. Wahidabanu, "Lung tumor detection and diagnosis in CT scan images," *2013 International Conference on Communication and Signal Processing*, Melmaruvathur, 2013, pp. 1108-1112.

[4] Prof. Anuradha S. Deshpande, Dhanesh D. Lokhande, "Lung Cancer Detection with fusion of CT and MRI images using image processing", International Journal of advanced research in computer engineering and technology (IRARCET), Vol.4 Issue 3, March 2015.

[5] S. Dabade, S. Chaudhari, S. Jadhav and A. Nichal, "A review paper on computer aided system for lung cancer detection," *2017 International Conference on Big Data, IoT and Data Science (BID)*, Pune, 2017, pp. 97-102.

[6] Shukla, Ayushi, ChinmayParab, Pratik Patil and Savita Sangam. "Lung Cancer Detection using Image Processing Techniques." (2018).

[7] J. Alam, S. Alam and A. Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier," *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, Rajshahi, 2018, pp. 1-4.

[8] K. Murphy, B. Ginneken, A. M. R. Schilham, B. J. Hoop, H. A. Gietema, and M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbor classification," Medical Image Analysis, vol. 13, issue 5, pp. 757–770, 2009.

[9] S. K. Kumar, J. Ramesh, P. T. Vanathi, K. Guunavathi, "Robust and automated lung nodule diagnosis from CT images based on fuzzy systems." in International Conference On Process Automation Control and Computing, Coimbatore, India, IEEE, 2011.

[10] Dr. LokaiahPullagura, Madhusudhan Rao Dontha, SrinivasaraoKakumanu, "Recognition of Fetal heart diseases through machine learning techniques", Annals of the Romanian Society for Cell Biology., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 2601-2615.

[11] LokaiahPullagura, JeevaKatiravan, "Train Delay Prediction Using Machine Learning", International Journal of Engineering and Advanced Technology, vol 9 Issue-2, December 2019.

[12] P. Lokaiah, S. Nyamathulla, M. Kiran kumar, KBV Rama Narasimham, "Performance Evaluation of Different Machine Learning Techniques for Prediction of Diabetes", Journal of Critical Reviews, vol 7, Issue 18, 2020.

[13] Dr. P. RatnaBabu, P. Lokaiah, "An effective noise reduction technique for class imbalance classification", International Journal of Psychosocial Rehabilitation, vol 24, Issue 04, 2020.

[14] LokaiahPullagura, Dr. Anil kumar, "Analysis of Train Delay Prediction System based on Hybrid Model", Journal of Advanced Research in Dynamical & Control Systems, Vol. 12, 07-Special Issue, 2020.

[15] LokaiahPullagura, Dr. Anil Kumar, "An Effective LSTM Network Model For Accurate Prediction of Delays in Indian Railway Networks", International Journal of Advanced Trends in Computer Science and Engineering, Vol 9, No.4, July-August 2020.