Lung Cancer Detection by Classifying CT Scan Images using Gray Level Co-Occurrence Matrix (GLCM) and K-Nearest Neighbours

Aayush Kamdar[1], Vihaan Sharma[1], Sagar Sonawane[1], and Nikita Patil[1]
[1]Department of Computer Engineering, Atharva College of Engineering
ashbkamdar@gmail.com
vihaansharma62@gmail.com
sonawanesagar2106@gmail.com
nikitapatil@atharvacoe.ac.in

Abstract. Lung cancer is the unbridled growth of abnormal cells in the lungs, as the growth of these abnormal cells continues tumors are formed which interfere with the normal functioning of the lung. Early cancer diagnoses, combined with treatment and proper medical care, enhances survival and cure rates. This study of Lung Cancer Detection has divided into four stages: a pre-processing stage, image enhancement stage, feature extraction stage, and cancer classification stage. The system thoroughly focuses on detecting Lung cancer disease with various image processing and machine learning techniques. Computerized Tomography (CT) Scan is an imaging technique used in the system to diagnose lung cancer. This study aims to classify lung cancer into benign and malignant lung cancer using CT scan images. The testing of this system on the given dataset of lung CT scan images, the system has shown a classification accuracy of 92.37% for determining benign or malignant cancer.

Keywords: Lung cancer detection, CT scan images, image processing, machine learning, feature extraction, cancer classification.

1 Introduction

Cancer is often thought as an untreatable, unbearably painful disease with no cure. India is likely to have over 17.3 lakh new cases of cancer and over 8.8 lakh deaths due to the disease by 2020 [9]. The National Cancer Registry Programme Report 2020 estimates there will be 13.9 lakh cases of cancer in India in 2020 [11], and that this number is likely to rise to 15.7 lakh by 2025. Lung cancer is known to spread fast even to the neighbouring organs. Lung cancer consists of two main types: Small Cell Lung Cancer (SCLC) and Non-small Cell Lung Cancer (NSCLC). Lung cancer tumors can be classified into two types: benign tumors (non-cancerous) and malignant tumors (cancerous). Lung Cancer in India constitutes 6.9% new cancer cases, also makes up for 9.3% of all cancer deaths in both genders [10]. It is one of the major cancers causes of death.

However, it is a misconception to think that all forms of cancer are untreatable and deadly. Our goal is to design a system where through image processing and machine learning technologies, the system can help the patient diagnose if the patient has benign tumor or malignant tumor. Early treatment steps can have a big advantage in increasing the survival rate of the patient.

Computerized tomography (CT) scans is one of the standard and most used imaging detecting medical techniques used for detection and diagnosing of lung cancer. These CT scan images are digitally stored in the DICOM format (.dcm). Lung CT scans are used as input of the system and are converted into grayscale images.

In the previous study "Lung Cancer Detection Based On CT-Scan Images With Detection Features Using Gray Level Co-Occurrence Matrix (GLCM) and Support Vector Machine (SVM) Methods ", they used GLCM which is considered as one of the highly efficient image feature extraction technique. For classification they have used Support Vector Machine (SVM) which is not as efficient as kNN classification (Mohd Firdaus Abdullah, 2020). In addition, we have implemented a Gabor filter for image enhancement which will deduct the noise in the input image. The enhanced images by Gabor filter are further taken as input for feature extraction by using Gray Level Co-Occurrence Matrix (GLCM) to get statistical data about the image which will be used for kNN classification. The classification stage will classify if the lung tumor is benign or malignant based on statistical data of the training dataset.

2 Literature Survey

In the research conducted by Qurina Firdaus, RiyantoSigit, et al., entitled "Lung Cancer Detection Based On CT-Scan Images With Detection Features Using Gray Level CoOccurrence Matrix (GLCM) and Support Vector Machine (SVM) Methods" [1], they implement the process of lung cancer detection using GLCM and SVM methods where they conduct a six stage process which includes Input CT scan images, Pre-processing images, Segmentation, Feature Extraction, Classification and Output based on whether the cancer is normal, benign or malignant. They make use of different methods in each stage to get accurate results to detect cancer. This research implementation of detecting lung cancer as benign or malignant has an accuracy level of 83.33%.

Another research conducted by Mohd Firdaus Abdullah, Siti Noraini Sulaiman, et al., entitled "Classification of Lung Cancer Stages from CT Scan Images Using Image Processing and K-nearest Neighbours" [2] implement the same kind of process but using image processing and classification by kNN method. After the analysis and implementation of this research, they found out that the kNN method has a higher accuracy.

In the research conducted by Sanjukta Rani Jena, Dr. Thomas George and Dr.NarainPonraj entitled "Texture Analysis based Feature Extraction and Classification of Lung Cancer" [3]. The paper covers three types of feature extraction which are - Shape-based FETs, Texture based FETs, Intensity-based FETs.

In the research paper conducted by A.P AyshathThabsheera and T.M Thasleema titled "Lung Cancer Detection Using CT Scan Images: A Review on various Image Processing Techniques" [4], they have conducted a review on various techniques of image enhancing and image classification used for detection of Lung Cancer. The paper reviewed many image enhancement techniques such as Gabor Filter, Median Filter etc. They further review some segmentation methods such as thresholding, Sobel edge detection etc. Also, feature extraction was summarized and a few of the features extracted were explained.

In the research conducted by Kusworo Adi, Catur Edi Widodo, Aris Puji Widodo, Rahmat Gernowo, Adi Pamungkas, and Rizky Ayomi Syifa titled as "Detection Lung Cancer Using Gray Level Co-Occurrence Matrix (GLCM) and Back Propagation Neural Network Classification" [5] implement the usage of GLCM feature extraction method. The research makes use of GLCM features like contrast, correlation, energy, and homogeneity to extract the data from the input. These features along with entropy and dissimilarity will be utilized in our research paper.

In another research conducted by Avinash S., Dr K. Manjunath, Dr S. Senthil kumar titled as "An Improved Image Processing Analysis for the Detection of Lung Cancer using Gabor Filters and Watershed Segmentation Technique" [6] implements image processing techniques like Gabor Filter for image enhancement and Watershed segmentation for image segmentation. This aim of this study proves that this new technique with the use of Gabor filters and watershed segmentation can be used for quick detection of lung cancer.

In a research conducted by Khin Mya Mya Tun, Aung Soe Khaing titled as "Feature Extraction and Classification of Lung Cancer Nodule using Image Processing Techniques" [7]. The authors utilize Median filter for image preprocessing which is followed by Otsu's thresholding for the segmentation of image. GLCM method is used for feature extraction and the classification of the images are done by Artificial Neural Network (ANN). The results of this study show that this system can be very important for early diagnosis of the disease while being cost and time efficient.

3 Proposed Methodology

3.1 Problem Definition

Cancer is a catastrophic disease which sets panic into people as there is a lack of a clear way to how to deal with cancer. Firstly, the majority of the people don't know about the early symptoms of various cancer diagnoses and as a result, tend to ignore the symptoms following their cancer getting worse. Prognosis of cancer survival rates when the cancer is detected in advanced stages is poor. Patient reports are hard to maintain because of various tests, diagnostics and lots of past medical documents. The current system is flawed in such a way that the doctors spend more time making decisions due to the cure for cancer not being definitive. Diagnoses of cancer by the current methods are inefficient as existing standard procedures are time-consuming and error-prone. In a world where Artificial Intelligence is becoming integrated with everything, medicine should be at the forefront of it. There are very few methods of predicting (determining) diagnosis for various diseases with the help of AI and Machine Learning techniques.

3.2 Proposed System

The system will take an input in the form of lung CT scan images and then follow the four stages which are pre-processing stage, image enhancement and segmentation stage, feature extraction stage and image classification stage which is described below in Figure 1.

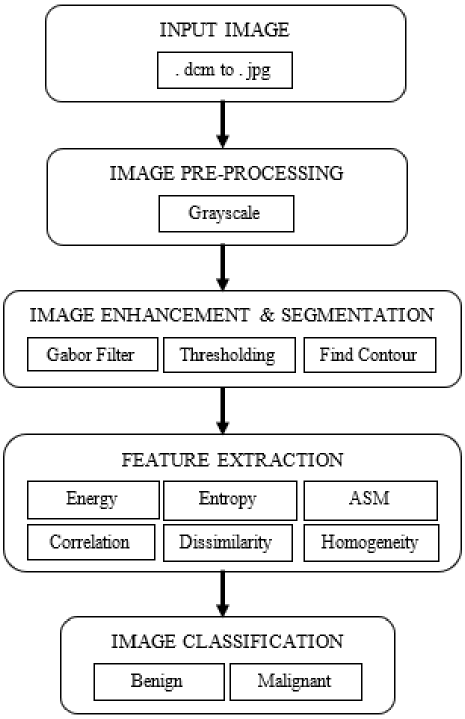

Figure 1. Proposed System

3.2.1 Input Image

A computerized tomography (CT) scan is an imaging technique used in the medical department to get cross-sectional (slices) images of the focused body part. CT scan uses a variety X-ray images taken from around the focused body part and combines them. The system takes a CT scan image input in the form of DICOM image, which is converted into JPEG image to help perform the research as it is easier to study and analyze the image better. Figure 2 shows an example of lung CT scan image which is used in this study.

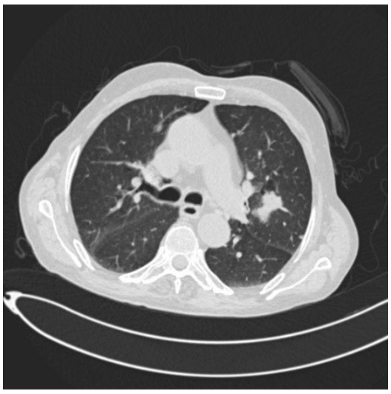

Figure 2. Lung CT Scan Image

3.2.2 Pre-Processing

The input image is now converted into a grayscale image to ease up the further process. We convert the image to grayscale in order to simplify the complexity of image and to extract the intensity data. Also, its faster to process grayscale image rather than any coloured image.

3.2.3 Image Enhancement

The next stage is Image enhancement, which is a process of improving the quality and information content of original data before processing. We use Gabor filter which is an adaptive filtering technique, which uses bandpass filters for applications like texture analysis or feature extraction. These filters apply a certain kind of band frequencies to an image, and also apply various gabor parameters functions (1) like sigma, lambda, gamma and phi into what we call a gabor kernel. This kernel helps in processing the gabor filter, which spread across the function in two dimensions according to what we pass to the values of the parameter, giving a filtered image capable read the textures and to extract features of the image.

$$G (x,y; \lambda,\theta,\psi,\sigma,\gamma) = \exp (- (x'^2+\gamma^2 y'^2) / 2\sigma^2) \exp (i (2\pi (x' / \lambda) + \psi)) \qquad (1)$$

After this we use Image thresholding which is one of the good segmentation technique used for images with significant differences in intensity values between the background and the main object, to separate the desired object from the background. This technique is used to obtain areas that contain cancers and convert grayscale images into binary images i.e., simply black and white.

3.2.4 Feature Extraction

After enhancing the images, these images are passed into feature extraction stage. Feature extraction is a technique to extract vital features like energy, entropy, contrast, etc are extracted which will help in the further stage for classification of cancer. Gray Level Co-Occurrence Matrix (GLCM) is used for feature extraction on the enhanced images. A Co-Occurrnce matrix is a matrix that is defined over an image to be the distribution of co-occurring pixel values at a given offset. Basically it is a texture filter functions which provide a statistical view of texture based on the image histogram. Here the offset distance is one pixel and the angles are 0, pi/4, 3pi/6, pi/2. The properties used are energy(2), correlation(3), dissimilarity(4), homogenity(5), ASM(6) and entropy(7).

$$\text{Energy: } \sum_{i,j} p(i,j)^2 \qquad (2)$$

$$\text{Correlation: } \sum_{i,j} \frac{(i - \mu i)\ (j - \mu j)\ p\ (i,\ \ j)}{\sigma_j\ \ \sigma_i} \qquad (3)$$

$$\text{Dissimilarity: } \sum_{i,j=0}^{N-1} P_{i,j}\ |\ i - j\ | \qquad (4)$$

$$\text{Homogeneity: } \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2} \qquad (5)$$

$$\text{ASM: } \sum_{i,j=0}^{N-1} P^2{}_{i,j} \qquad (6)$$

$$\text{Entropy: } \sum_{i,j=0}^{N-1} -1n\ (P_{ij}) P_{ij} \qquad (7)$$

The statical data or extracted features thus obtained on an image are saved in a form of .csv file. The data saved in the csv file is used for analysis and image classifcation.

3.2.5 Image Classification

The final stage is image classification stage where the lung cancer image is classfied into benign or malignant. This will be acheived by using kNN machine learning model. The csv files generated in the past stage is passed onto kNN machine learning model for the classification of lung cancer. kNN or k-nearest neighbours is a supervised machine learning model which assumes the similarity between the new case data and available cases, and put the new case into the category that is most similar to the available categories. First the training dataset is passed to train the model then the test dataset is passed on the trained model to predict or classify the cancer.
Applying kNN model to the test cases data, the input is the parameter values generated from the feature extraction stage. These parameter values will be matched with the parameter values of the train cases data stored in the database with a .csv file. The output of this process is 0 or 1, where 0 is benign lung cancer and 1 is malignant lung cancer.

4 Experimental Results and Analysis

CT Scan images that are taken as input are first converted into grayscale as a pre-processing stage to study and analyze the image better.

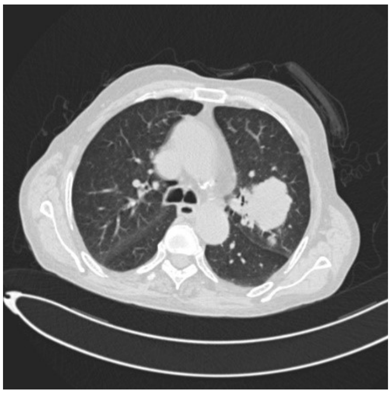

Figure 3. Grayscale Image

This graysacle images are then passed through the image enhancement stage which uses Gabor Filter. This filter removes all the noise in the image and clarifies the texture of the image to help extract the features. The output of the filtered image is given below in Figure 4.

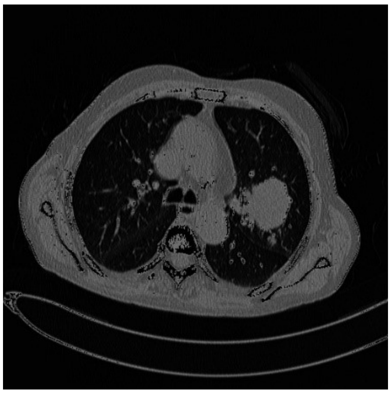

Figure 4. Gabor Filtered Image

After enhancing the image we move to image segmentation stage where we used thresholding technique to convert the image into binary image as shown in Figure 5. Segmentation partitions the image into multiple image segments. These segments defines the tumor which is meaningful and easier to analyze. As seen in Figure 6 where certain image parts are selected and we placed a contour over them.

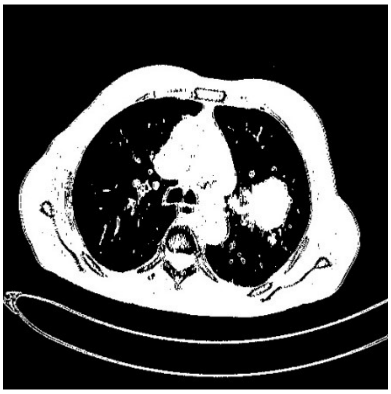

Figure 5. Thresholding

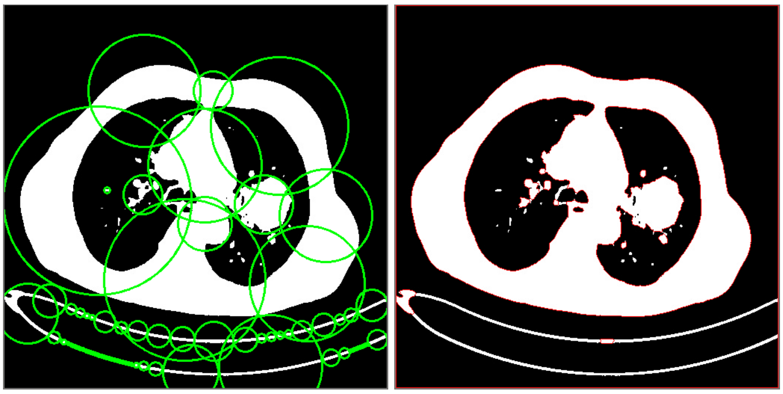

Figure 6. Contour Segmentation

The next step is feature extraction, where we find out different features for all the images, namely energy, entropy, homogeniety, dissimilarity, correlation and ASM (angular second moment). The features are extracted at every angle and at every offset of pixel range set by its parameters. These features are calculated using GLCM method and the values are stored in a .csv file. The following Table 1 shows different values of each features calculated by the GLCM method of different CT images.

T	Energy	Correlation	Dissimilarity	Homogeneity	ASM	Entropy
0	0.181989	0.737039	12.62417	0.319139	0.03312	4.90813
0	0.146102	0.7344	11.85163	0.286115	0.21346	4.920787
1	0.13742	0.875047	7.780956	0.361196	0.01888	5.009644
1	0.139809	0.885935	7.47172	0.36514	0.01955	5.013208

Table 1. Features extracted for Benign cases (0) and Malignant cases (1)

For classification of the lung cancer, we first train the model on above extracted features on train dataset. And then, on passing the test dataset on the trained kNN model, the system is getting an accuracy of 92.37% on the test dataset. The study concludes that the accuracy can be increased even further if improved image segmentation process is used.

5 Conclusion

Knowledge of early symptoms and diagnosis of cancer is vital towards the prevention and cure for cancer. Thus, to be able to provide help in the treatment of cancer we have decided to improve pre-existing methods and to make a system that can be of great use in the treatment process. This study will help hospitals by detecting cancer through image processing which will provide better results. The system takes a minimal amount of time to complete the detection of cancer which may help the doctors to provide better treatment care. We have combined image enhancement, feature extraction, and machine learning methods which shows an accuracy of 92.37% on the test dataset.

References

1. Q. Firdaus, R. Sigit, T. Harsonoand A. Anwar, "Lung Cancer Detection Based OnCTScan Images With Detection Features Using Gray Level Co-Occurrence Matrix (GLCM) and Support Vector Machine (SVM) Methods," 2020 International Electronics Symposium (IES), Surabaya, Indonesia, 2020, pp. 643-648, DOI: 10.1109/IES50839.2020.9231663
2. M. Firdaus Abdullah, S. Noraini Sulaiman, M. Khusairi Osman, N. K. A. Karim, I. Lutfi Shuaib and M. Danial Irfan Alhamdu, "Classification of Lung Cancer Stages from CT Scan Images Using Image Processing and k-Nearest Neighbours," 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC), Shah Alam, Malaysia, 2020, pp. 68-72, DOI: 10.1109/ICSGRC49013.2020.9232492.
3. S. R. Jena, T. George and N. Ponraj, "Texture Analysis Based Feature Extraction and Classification of Lung Cancer," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2019, pp. 1-5, DOI: 10.1109/ICECCT.2019.8869369.
4. AyshathThabsheera A.P., Thasleema T.M., Rajesh R. (2019) Lung Cancer DetectionUsing CT Scan Images: A Review on Various Image Processing Techniques. In: Nagabhushan P., Guru D., Shekar B., Kumar Y. (eds) Data Analytics and Learning. LectureNotes in Networks and Systems, vol 43. Springer, Singapore. https://doi.org/10.1007/978981-13-2514-4_34.
5. Adi, Kusworo & Widodo, Catur & Widodo, Aris & Gernowo, Rahmat & Pamungkas, Adi & Syifa, Rizky. (2018). Detection Lung Cancer Using Gray Level Co-Occurrence Matrix (GLCM) and Back Propagation Neural Network Classification. Journal of Engineering Science and Technology Review. 11. 8-12. 10.25103/jestr.112.02.
6. S. Avinash, K. Manjunath and S. S. Kumar, "An improved image processing analysis for the detection of lung cancer using Gabor filters and watershed segmentation technique," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2016, pp. 1-6, doi: 10.1109/INVENTIVE.2016.7830084.
7. Khin Mya Mya Tun and Aung Soe Khaing, "Feature Extraction and Classification of Lung Cancer Nodule using Image Processing Techniques", International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 3, March 2014.
8. M. Šarić, M. Russo, M. Stella and M. Sikora, "CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images," 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 2019, pp. 1-4, doi: 10.23919/SpliTech.2019.8783041.
9. Explained: The Cancer crisis in India, `https://indianexpress.com/article/explained/cancer-crisis-in-india-national-cancer-institute-5598853/`
10. Lung Cancer: Prevalent Trends and Emerging Concepts, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4405940/`

11. Healthworld – Economic Times, https://health.economictimes.indiatimes.com/news/diagnostics/india-to-have-13-9-lakh-cancer-cases-by-year-end-15-7-lakh-by-2025-icmr/78572244
12. National Foundation for Cancer Research, https://www.nfcr.org/cancer-types/lung-cancer/