Content Based Image Retrieval System for Lung Cancer Detection Using Neural Network and Circular Hough Transform

Rejiram R\textsuperscript{1}, Kanniga E\textsuperscript{2}

\textsuperscript{1}Research Scholar, Department of Electronics and Communication Engineering, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India
\textsuperscript{2}Department of Electronics and Communication Engineering, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India

Article History:
Received on: 26 Oct 2020
Revised on: 27 Nov 2020
Accepted on: 01 Dec 2020

Keywords:
PNN, CAD, LIDC, Circular Hough Transform, Benign, Malignant

ABSTRACT

Computer Aided Detection (CAD) systems that automatic detection and localize lung nodules in CT scans. A major problem in this system is a large number of false positives because of no provision for comparison of the predicted output. This paper recommends a new system with a combination of CBIR and neural network to full fill the gap in the area of early detection of lung cancer. From the preprocessed CT scan image, the system identifies whether it contains nodules using Circular Hough Transform and classifies into benign or malignant nodule using Probabilistic Neural Network. Then, it searched for the most identical pictures and retrieved it from the database. From the retrieved image, it is easy to identify the present cancer stage of the patient. Experiments have done based on both LIDC database and the locally collected database. The performance evaluation of the system is done by using both. The experimental results show that the present study easily differentiates benign and malignant nodules with an efficiency of 97% accuracy on LIDC dataset, 95% accuracy on Local dataset and similar images are retrieved with its present stage from the available database with a higher precision and recall rate.

*Corresponding Author
Name: Rejiram R
Phone: +91- 9495124360
Email: rejiramnair@gmail.com

ISSN: 0975-7538
DOI: https://doi.org/10.26452/ijrps.v11i4.3957

INTRODUCTION

Lung cancer is ranked second as the most common form of cancer around the world. Radiologists nowadays use a computed tomography (CT) scans of the chest to detect lung tumors as it has high sensitivity and low error rate. However, even with these CT scans, it requires a certain amount of time and experience to detect and label the lung tumors. Moreover, considering the large number of cases that radiologists have to analyses on a daily basis, they will experience pressure and a heavy workload. A solution to reduce this pressure and workload is to use computer-aided detection (CAD) system that uses automatic detection, which localizes lung nodules in CT images. These systems are helpful to assist the radiologists in the process of lung tumors detection. They have many benefits, such as reducing the error rate of nodule detection, reducing the operation time and detecting tumors that are overlooked by the radiologists. Several studies have shown that CAD systems offer a useful second opinion. But, one of the problems in this system is a large number of false-positive which may occur during detection. To solve this, content-based medical image retrieval system is used. Firstly database
prepared as benign and malignant using neural network detection. Then, when a query image is fed to the system, it classifies it into benign and malignant and then it retrieves the most identical image from the database along with the cancer stage. This system can aid the medical field a lot.

In (Dandl, 2018) a recent paper, the nodule detection and classification were done simultaneously. Another advantage is based-on SOM method during segmentation for the detection of small nodules in the lung. The model is flexible, which can handle the variation of shape and size of nodules, which is used to adjust the labels class in proper (Sori et al., 2019). Region growing algorithms are implemented to detect nodules in lungs from a CT Scan image of Lungs, which suggest CT based pulmonary parenchyma detection for getting an accurate and effective result (Chaudhary and Singh, 2012). With the selected size of the structural element, identify the large size juxta-pleural nodules and are eliminated, so it is suitable for isolated solid nodules detection. This can be executed in a better way in the proposed method (John and Mini, 2016).

Here we describe a scheme for diagnosing based on the computer-aided methodology for lung nodules detection and to propose a method to evaluate the similarity between the reference lung nodule dataset and query lung nodule. At the initial stage, preprocessed CT image fed to the system, Lung mask extracted using LUVEM algorithm, identifies nodules using Circular Hough Transform [CHT]. Then, three groups of features are extracted from the database based on these nodules are classified into benign and malignant using probabilistic neural network. Lastly, a CBIR approach proposed to retrieve similar images which are closely similar to the query, from the database and finally accuracy in the classification used as the performance assessment index. The results in the performance evaluation stage show that the accuracy of the proposed system is more than 97% and 95% in the standard database and the locally collected database, respectively. The images are retrieved with higher precision and recall rate.

Proposed work

The proposed system consists of a chain of works which had the data acquisition, preprocessing, Lung volume extraction, nodule detection after the segmentation, nodule classification and finally content-based image retrieval as shown in Figure 1.

Image Acquisition

In the image preprocessing stage, that is the initial stage of the proposed work in which reading
**Table 1: GLCM Features Comparison**

| GLCM Features     | Benign     | Malignant  |
|-------------------|------------|------------|
| Contrast          | 1.04997    | 1.22824    |
| Correlation       | 0.00602    | 0.013074   |
| Cluster Prominence| 0.87967    | 0.961336   |
| Cluster Shade     | 0.87967    | 0.961336   |
| Dissimilarity     | 1.51303    | 33.59047   |
| Energy            | 0.35612    | 0.693887   |
| Entropy           | 0.00387    | 0.004838   |
| Homogeneity       | 0.98170    | 0.981364   |
| Homogeneity       | 0.06901    | 0.076555   |
| Contrast          | 0.99833    | 0.998292   |

**Table 2: Different features accuracy on LIDC database and Local database**

| Features Used | Local Database | LIDC database |
|---------------|----------------|---------------|
| SSF           | 78%            | 81%           |
| GLCM          | 92%            | 93%           |
| Wavelet       | 90%            | 91%           |
| Combined      | 95%            | 97%           |

**Table 3: Performance Evaluation of PNN Neural Network Using Confusion Matrix**

| N=100          | Benign | Malignant | Total |
|----------------|--------|-----------|-------|
| Benign         | TN = 47| FP = 3    | 50    |
| Malignant      | FN = 0 | TP = 50   | 50    |

**Table 4: Performance evaluation of CBIR (Benign)**

| Number of Images | Number of True Retrievals | Precision Rate |
|------------------|---------------------------|----------------|
| 5                | 5                         | 100%           |
| 10               | 10                        | 100%           |
| 15               | 14                        | 93%            |
| 20               | 18                        | 90%            |

**Table 5: Performance evaluation of CBIR (Malignant)**

| Number of Images | Number of True Retrievals | Precision Rate |
|------------------|---------------------------|----------------|
| 5                | 5                         | 100%           |
| 10               | 10                        | 100%           |
| 15               | 15                        | 100%           |
| 20               | 19                        | 95%            |
of CT images is performed from the LIDC database, which is stored as DICOM (Digital Imaging and Communications in Medicine) (McNitt-Gray et al., 2007) format in the LIDC dataset and locally collected database. These images are converted into JPEG format. The pixel size of a converted image is in 512 x 512 with JPEG format.

Preprocessing

The goal of image preprocessing step is mainly concentrating on preventing misleading results which may occur in the subsequent processes. Here, the sharpening of the input done. Identifying the contrast between the various colors is considered as sharpness. A quick transition from white to black and vice versa are looked sharp. When the transition is gradual from black to gray to white looks as blurry. The contrast along the edges in the various shapes in the image, is clear with different colors meet, is consider as sharpened image. Sharpening returns an enhanced version of the true color or grayscale input image where the edges, a common image features, have been sharpened using the unsharp masking method. A common image sharpening technique in digital image processing is the unsharp masking (USM).

Lung Volume Extraction

By considering the image processing based morphological operations, the Lung lobes are extracted from the obtained CT images. The unrelated segments on the edges and sides of the image can be removed using the LUVE algorithm, at the time of pre-processing.

Steps of Lung Volume Extraction

1. Creates an initial lung mask from the sharpened image
2. Find RGB values used for thresholding of marker/color image.
3. User select multiple pixels on the marker image
4. Creating an output image based on selected pixels on the marker
5. Morphological operation Filling and dilation are done.

Lung Volume Segmentation

The nodule segmentation is the next step of nodule detection. The nodule segmentation is done using the K-means clustering algorithm.

The following are the steps of K means clustering algorithm,

1. Initially Place a K points into space, which represented by the objects that are being clustered. And these points considered as initial group centroids.
2. The object in the closet centroid is considered as a group.
3. When all objects have been assigned, the positions of the K centroids are recalculated.
4. The method of recalculating the new cluster center by using the formula: \( V_i = \frac{1}{a} \sum_{j=1}^{a} x_i \) where a represents the number of data points in the \( i^{th} \) cluster.
5. Steps 2 and 3 are repeated until the centroids no longer move.
6. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Nodule Detection

In this stage, the location of the pulmonary nodule is determined. Lung nodules are to be detected before nodule segmentation. Mostly, nodules are in a helical and circular structure. Hence, it can be detected using circularity property. Many methods have been used commonly for identifying rounder objects present in the images. Circular Hough Transform is one of the commonly used famous methods.
**Figure 8:** CBIR in LIDC Database (Benign)

**Figure 9:** Malignant data from LIDC database with CBIR
for detection of rounder objects in the image (Duda and Hart, 1972). In order to detect the nodules in various size different radiuses are chosen in doing Circular Hough Transform. One of the famous methods used in Digital Image Processing for detecting circular objects in a digital image is the Circular Hough Transform (CHT). In this feature extraction technique, we can detect circles. It is derived from the Hough Transform. The main purpose of this technique is to find circles in imperfect image inputs. By voting in the Hough parameter in the space and then select the local maxima, the circle candidates are detected, the obtained values are called accumulator matrix.

The CHT relies on equations or circles (Duda and Hart, 1972). \( r^2 = (x - a)^2 + (y - b)^2 \) indicates the equation of the circle with center at \( x, y \) and radius \( r \). The Hough transform used to identify the parameters of a circle with the number of points that fall on the perimeter are known. The parametric equation of the circle with center \( (a, b) \) and radius \( R \) can be described as, \( y = b + R \sin \theta \).

**Feature Extraction**

From the segmented nodules in the image, various features which are required for the classification are extracted. Based on these extracted features, the classification of nodule into benign and malignant is done. Firstly, from the gray level histogram values of the segmented nodule image, the first-order statistical features were calculated.

In this study, six basic features such as means, standard deviation, skewness, entropy, variance and kurtosis are calculated (Mao et al., 2000). GLCM features are also extracted from the image.

A total of 88 GLCM features were extracted within the various angles directions like \( 0^\circ \), \( 45^\circ \), \( 90^\circ \), and \( 135^\circ \) with distance, \( d=2 \). Wavelet decomposition transform denotes the different region (TEF) based energy distribution feature. In 2D wavelet decomposition, ROI (Region of Interest) of the CT image is divided into four sub-bands.

From the processed image, obtain three images with low frequencies and one with high frequencies with the help of wavelet decomposition transform. Extracted 2D Features from CT Image of Lungs are 6 statistical features, 88(22x4), GLCM features and 13 Wavelet features.

**Nodule Classification**

The Probabilistic Neural Network (PNN) is used here for classification. According to the number of input 107, Neuron number are selected. PNN is the simplest neural network and takes less time for classification (Mao et al., 2000).

**CBIR (Content Based Image Retrieval)**

The most similar images are retrieved from the LIDC database using CBIR technique. The images are stored as malignant and benign as separately along with its medical description. When an input is fed to the system, the system classifies into benign and malignant and secondly it retrieves most similar images from the database using Euclidean distance measurement. In this work, Euclidean distance is used for image retrieval. It is known as L2 distance. If \( u = (a, b) \) and \( v = (c, d) \) are two points. The Euclidean distance between \( u \) and \( v \) is given by: 
\[
(u, v) = \sqrt{(a - c)^2 + (b - d)^2}
\]

Based on these measurements, images are retrieved in the order of similarity.

**RESULTS AND ANALYSIS**

The proposed system was implemented using MATLAB. The dataset for the project was collected from LIDC dataset and from the local hospital for validating the results. A total of 100 images were taken from the LIDC database out of that 50 were benign, and 50 images were malignant. For testing the performance of the system, 10 images that belong to two different classes were used. After that validation was done using 60 images from the local database. Initially select a sample image from the database in Figure 2.

The sample images are then sharpened for removing the noises occurred during imaging of lungs using CT scan, as shown in Figure 3. The lung volume is extracted after the morphological operation is shown in Figure 4. The lung volume is extracted after the morphological operation is shown in Figure 4. The lung volume is extracted after the morphological operation is shown in Figure 4. The lung segmentation is done through K-mean clustering is shown in Figure 5. The lung nodules are detected using Circular Hough Transform with smaller and larger radii is as shown in Figure 6. The lung nodules with smaller and larger radii are extracted, which is shown in Figure 7.

From the extracted nodules in each of benign and malignant, six statistical features, 22 GLCM features and wavelet features are extracted. Out of that more accurate results obtained from GLCM features and more weightage is given to that. Finally, features fed to the neural network and classification is done. In Table 1, GLCM features of benign and malignant nodules are given.

Most of the GLCM features have high values for malignant nodules. For getting accurate results, these features are fed to the neural network. In Table 2, shows that the LIDC database has higher accuracy on neural network detection of cancer. The Performance evaluation of PNN neural network
is done using Confusion matrix is shown in Table 3. The content-based image retrieval of a benign and malignant tumor are shown in Figure 8 and Figure 9. The performance evaluation of CBIR is shown in Table 4 and Table 5.

CONCLUSION

The proposed Lung cancer detection system implemented successfully with 97 % accuracy using PNN neural network. Lung nodule database prepared as benign and malignant. From the set of large unsorted LIDC database, similar images are retrieved using content-based image retrieval, (CBIR). The performance of the proposed system compared with the local database also for validation and obtained an accuracy of 95% detection. This system helps the doctor to improve the treatment level by correct diagnosis and also help to identify the present stage of the patient with the help of retrieval from a large database. In the future, successful treatment method of each patient can be attached with the image which can act as a reference.

Conflict of Interest

The authors declare that they have no conflict of interest for this study.

Funding Support

The authors declare that they have no funding support for this study.

REFERENCES

Chaudhary, A., Singh, S. S. 2012. Lung Cancer Detection on CT Images by Using Image Processing. International Conference on Computing Sciences, pages 142–146.

Dandil, E. 2018. A Computer-Aided Pipeline for Automatic Lung Cancer Classification on Computed Tomography Scans. Journal of Healthcare Engineering, 2018:1–12.

Duda, R. O., Hart, P. E. 1972. Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM, 15(1):11–15.

John, J., Mini, M. G. 2016. Multilevel Thresholding Based Segmentation and Feature Extraction for Pulmonary Nodule Detection. Procedia Technology, 24:957–963.

Mao, K. Z., Tan, K. C., Ser, W. 2000. Probabilistic neural-network structure determination for pattern classification. IEEE Transactions on Neural Networks, 11(4):1009–1016.

McNitt-Gray, M. F., Armato, S. G., Meyer, C. R., et al. 2007. The Lung Image Database Consortium (LIDC) Data Collection Process for Nodule Detection and Annotation. Academic Radiology, 14(12):1464–1474.

Sori, W. J., Feng, J., Liu, S. 2019. Multi-path convolutional neural network for lung cancer detection. Multidimensional Systems and Signal Processing, 30(4):1749–1768.