Lung Cancer Detection using Convolutional Neural Network

Sayana Sharma, Mandeep Kaur, Deepak Saini

Abstract: The mortality rate is increasing among the growing population and one of the leading causes is lung cancer. Early diagnosis is required to decrease the number of deaths and increase the survival rate of lung cancer patients. With the advancements in the medical field and its technologies CAD system has played a significant role to detect the early symptoms in the patients which cannot be carried out manually without any error in it. CAD is detection system which has combined the machine learning algorithms with image processing using computer vision. In this research a novel approach to CAD system is presented to detect lung cancer using image processing techniques and classifying the detected nodules by CNN approach. The proposed method has taken CT scan image as input image and different image processing techniques such as histogram equalization, segmentation, morphological operations and feature extraction have been performed on it. A CNN based classifier is trained to classify the nodules as cancerous or non-cancerous. The performance of the system is evaluated in the terms of sensitivity, specificity and accuracy.

Index Terms: CAD, CNN, GLCM, Image processing, Lung cancer, Threshold Segmentation.

I. INTRODUCTION

Lung cancer is the second most common cancer in both men and women. About 13% of new cancers are lung cancer. According to the estimate of American cancer society for 2019 the estimated new cases are about 228,150 of lung cancer in which 116,440 are men and 111,710 are women and estimated deaths are about 142,670 due to lung cancer in which 76,650 are men and 66,020 are women. The survival rate of people suffering from this disease depends on which stage the cancer is been diagnosed. Early stage cancer can be cured which may decrease the growing rate of deaths due to lung cancer. In the recent years, the people showing the symptoms of lung cancer can be detected by a test known LDCT (low dose-computed tomography) or CT scan of the chest which have helped the doctors or radiologists to analyze the shape, size and position of lung nodule or tumor which have spread across the body [1]. CT scan images are used for the detection of lung or pulmonary (means related to lungs) nodules because of its high sensitivity. A pulmonary nodule is an irregular, small oval shaped or round figured nodule with a diameter up to 30mm in the lung [2]. If a pulmonary nodule is present in the lungs it indicates the cancer is at an early stage.

CT scanner takes many detailed pictures or images of the internal organ which are combined by computer into the image of slices of part of the body in which few nodules or the non-pathological structures which are round figured in the shape but, all are not exactly cancerous. So it becomes very time consuming and tedious task to analyze the tiny or small nodules for a radiologist [3]. However radiologists can be assisted by CAD (computer aided diagnosis or detection) system. CAD system processes the digital images to highlight the conspicuous area or section affected by the disease. It helps doctors and radiologists to interpret the medical images. This technology has combined artificial intelligence, machine learning algorithms and computer vision with image processing. The CAD system can examine large number of CT images using image processing techniques such as preprocessing, segmentation, feature extraction, image acquisition and also deep learning based methods for better and early diagnosis of pulmonary nodules [4]. This research work presents a novel approach to the CAD system based on CNN (convolutional neural network) and different image processing techniques [5, 6, 7]. The approach proposed a method where CT scan image is taken as input image on which different image processing techniques were applied step by step. 1) Preprocessing of the input image 2) Histogram equalization 3) Segmentation by thresholding technique 4) Morphological operations which included – image filtering, dilation, image filling. 5) Feature extraction. It is performed using GLCM algorithm to extract the features of the input image. Different textural features are extracted such as shape, contrast, entropy and many other features which helped in classification. Finally the classification of the detected nodules are done using CNN as classifier. This classification will classify whether the nodule is cancerous or non-cancerous. The system is trained and the parameters obtained are in the terms of sensitivity, accuracy, specificity and time taken by the system for each image to detect the cancer. Software used for the proposed method is implemented in MATLAB. The below given Fig. 1, shows the simple processing of the proposed CAD system. In this paper section II presents the literature survey and the proposed methodology is explained in the section III.

Revised Manuscript Received on August 25, 2019.
Sayana Sharma, currently pursing M.tech in Electronics and communication stream as final year student from Punjabi University, Patiala.
Mandeep Kaur, an Assistant Professor and is currently pursuing her Ph.D in Punjab University, Patiala.
Dr. Deepak Saini, Assistant Professor in Punjabi university, Patiala.
Lung Cancer Detection Using Convolutional Neural Network

Fig. 1 Block diagram for proposed CAD system

II. LITERATURE REVIEW

Amir Roointan, et.al reviewed the development of inclusive molecular description of tumor lump [8]. It was clarified that a fundamental role was played by the ailment biomarkers in the early detection and indulgent of tumor analysis. The presented work summarized the speedy development of biosensor equipment’s for lung tumor biomarkers discovery. In the near future, these intellect plans can have huge consequences on scheduled medical scrutiny of biomarkers of pharmacogenomics and pharmacogenomics importance. It was also noticed that more expansion in nano bio-techniques in association with nano bio-composite and miniaturization approaches would considerably improve existing bio diagnostic capability for sensing tumor biomarkers in genuine organic models with sufficient compassion, acuteness, sturdiness and price efficiency. Jing Songa, et.al proposed a novel approach of microscopic hyper spectral imaging for the identification of ALK affected lung tumor [9]. In this approach, a household microscopic hyper spectral imaging scheme was utilized for capturing the pictures of five classes of lung tissues. For the minimization of noticeable banding clamor and sound element from authentic information, a preprocessed algorithm was introduced. After this, a combination of support vector machine, mass study and clumping processing was used for the projection of a segmentation algorithm. The variation of ethereal graphs, virtual amount amid cytoplasm and cell core was utilized for the differentiation of the fluctuation of spectral curves and the relative proportion between cytoplasm and cell nucleus for the ALK affected and non-affected lung cancers. The tested outcomes demonstrated that ALK affected set contained 77.3% comparative amount of cytoplasm while the ALK positive set contained 40.6% cytoplasm comparative quantity. The investigational outcomes related to quantitative scrutiny and ethereal curves demonstrated that the treatment of ALK affected lung tumor implemented with low concentrated medicines would be developed towards the ALK non-affected lung tumor. Guobin Zhang, et.al presented a serious evaluation of the CADe scheme for automated lung cancer recognition with the help of CT descriptions for summarizing the existing developments [10]. In the initial stage, a brief description of CADe scheme was provided. After this, a comprehensive summary of the five main mechanisms in CADe scheme was offered. These mechanisms included information attainment, preprocessing, lung image segmentation, nodule recognition and false positive diminution. A brief summary of superior nodule detection methods and classifiers was also provided on the basis of understanding, false positive value and other constrained data. After different studies it was evaluated that CADe scheme was essential for timely lung malignancy recognition. Moritz Schwyzer, et.al estimated the usefulness of machine learning for lung tumor recognition in FDG-PET imaging in the scenario of ultralow amount PET scan [11]. In the absence of pulmonary tumor, the recital of artificial neural network on selective lung cancer patients was examined. The sensitivity rate of 95.9% and 91.5% was attained by the artificial neural system for lung cancer detection. The artificial neural network achieved precision of 98.1% and 94.2%, at average dosage and ultralow dosage PET 3.3%, correspondingly. The tested outcomes demonstrated that machine learning approach provided assistance to the completely automatic lung tumor recognition at extremely small and efficient radiation dosages of 0.11 mSv. It was also suggested that more advancements in this technique could enhance the accurateness of lung tumor testing approaches. Suren Makajua, et.al stated that CT images could be used for the lung tumor recognition. The major objective of this study was the evaluation of different automated technologies, investigation of existing finest method, recognition of its restrictions and disadvantages and the projection of a decisive system with several advancements [12]. For this purpose, the lung tumor recognition approaches were classified on the basis of their lung cancer analyzing accurateness. In every stage, these lung cancer recognition methods were examined and their restrictions and disadvantages were considered. It was identified that different lung cancer detection techniques showed different precision. Some techniques showed least precision rate while some techniques showed good precision rate for lung cancer detection but no technique showed 100% precise lung cancer detection. Madhura J, et.al presented a review of noise reduction approaches for lung cancer diagnosis [13]. It was stated that lung cancer was a solemn ailment which caused due to the abnormal growth of cells in the lung tissues. Amongst all the other kinds of tumors, the lung tumor was identified as the most incident cancer. Therefore this cancer became the reason of several cancer patients’ deaths. The early recognition of lung cancer was very important for protecting various lives. The presented review study demonstrated a brief overview of lung tumor. This review work also described the different kinds of noises present in the pictures, techniques for the attainment of apparent pictures and noise elimination methods. A brief review on the existing noise elimination methods was also provided in this paper.

III. RESEARCH METHODOLOGY

Working of CAD system in this research work starts with preprocessing of input CT scan image. Next histogram equalization, lung segmentation using thresholding technique, and different morphological operations are performed on the image. GLCM (gray-level co-occurrence matrix) algorithm is used for extraction of the features from the input image. In the last step classification is done using CNN classifier by training the neural network system and parameters are obtained in the result. The proposed methodology is explained step by step as given below:-

A. Preprocessing
The preprocessing is the first step of the system. In this CT scan image is taken as input image. The aim of this step is to convert the image into gray scale image and to remove the noise from them using a filter. Pre-processing can be done using different types of filters such as Gabor filter, median filter, Gaussian filter, wiener filter and erosion filter for removal of noise. The Gabor filter is used in this research to de-noise the image. It also increases the contrast of the image giving an enhanced version of original image. As we know that the medical images are grayscale by default but still the raw CT scan images taken as input images are still converted to grayscale to increase its quality. As shown in the below given Fig. 2, the input image after preprocessing in which contrast of the image is increased making the objects clear and more visible. And the noise is also removed.

**B. Histogram Equalization**

After preprocessing of input image histogram equalization is performed. It is a method which uses image’s histogram to adjust the contrast in image processing. It smoothen the histogram of an image and increases global contrast and adjustment of this contrast allows the better intensity distribution. The dark image represents the lower level histogram and overexposed image indicate higher level of histogram. Histogram of chest CT scan input image consists of peak and valleys representing different regions of the lungs which are equalized and a clear, crisper image with sharp borders and edges is obtained. It also highlights the required objects in the grayscale image giving enhanced version of the image.

**C. Segmentation**

In this step the image is converted into the binary image or black and white image as shown in the above Fig. 3(A). As the binary image is a digital image and its pixels can be represented in 0 and 1 which are two discrete levels. The level 1 indicates the presence of data that is white color and the level 0 indicates absence of the data that is black color. The goal of segmentation is to partition the digital image into multiple segments to locate the required objects so that it could be easy to analyze the nodules in the image. In this approach segmentation is done using Otsu’s thresholding technique. The segmentation of CT scan input image is done by removing those pixels which are below or above the constant level called discrete level (threshold value). In the presented work, a gray scale picture is utilized for thresholding process. The binary image attained from thresholding comprises several benefits like lesser storage space, speedy dispensation velocity and easiness in exploitation in comparison with gray level picture that generally includes 256 steps but binary image have two levels 0 and 1 making it easier to analyze for the radiologist the required information for detection of the nodules. As shown in the Fig. 3(D) segmentation will give the ROI (region of interest) that is the nodules which are conspicuous ones and needs to be diagnosed whether they are cancerous or not. ROI helps in feature extraction. Before feature extraction different morphological operations are applied which are explained in the next step.

![Fig. 2 Preprocessing of Chest CT scan image. (A) Raw CT image used as input image. (B) After the removal of noise –preprocessed image](image)

**D. Morphological Operations**

In this research the morphological operations are performed before feature extraction. In a morphological operation the adjustments of pixels in the images are done which is based on its neighbourhood pixel’s value. These operations are used to extract the image components which are useful in the description of region shapes and gives better representation. The operations used in this research are filtering, dilation and image filling. The operations applied are shown in the below given Fig. 4.

a) **Filtering** - This image filtering allows the morphological operators such as dilate, erosion open and close to be applied on the image with these filters. With image filtering the image regions can grow or shrink and can fill in or remove the image region boundary. The sobel edge detector is used for filtering in this research work.

b) **Dilation** - The addition of pixels to the boundaries of the object in the image is known as dilation. A structuring element added to the input image to get output image. The number of pixels added or removed depend on the shape and size of the element.

c) **Image Filling** - This operation fills the holes which can be defined as background region in the input binary image.
Lung Cancer Detection Using Convolutional Neural Network

Fig. 4 Morphological operations applied on the input image (A) Filtered image. (B) Dilated image. (C) Image filling.

E. Feature Extraction

Different features are extracted of the detected nodules in the image, using GLCM (Gray Level co-occurrence matrix) algorithm. The GLCM algorithm will extract the textural features of the input image. The gray co-occurrence matrix is created and several statistic are derived from the matrix which analyzes the textural features in the images. The GLCM algorithm extracts 13 features of the image for the tumor detection. These features allow to differentiate the nodules which are cancerous or not. Some of the features have been discussed below:

a) Energy - This feature is used for optimization or minimization. It performs gradient-descent and compute its lowest value and giving the desired output for image segmentation.

\[ \text{Energy} = \frac{\sum_{i=0}^{N-1} p_{ij}}{\sqrt{\sum_{i=0}^{N-1} p_{ij}}} \]  

(1)

b) Entropy - This feature is used to characterize the texture of an input image which describes measure of degree of the randomness.

\[ \text{Entropy} = \sum_{i=0}^{N-1} p_{ij} \log_{2} p_{ij} \]  

(2)

c) Contrast - Here contrast is the ratio of difference of maximum intensity value and minimum intensity value upon sum of maximum and minimum intensity value.

\[ \text{Contrast} = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} \]  

(3)

d) Mean - It is the summation of each neighbouring pixel when divided by its weight value.

\[ \text{Mean} = \frac{\sum_{w} \text{neighbouring pixels}}{w} \]  

(4)

e) Absolute mean - It can be defined as the absolute value of the mean which can be calculated as per given formula.

\[ X = \sum_{w} \text{abs (neighbouring pixels)} / w \]  

(5)

f) Standard Deviation - It is the square root of the average squared deviation of each pixel from its mean.

\[ \sigma = \sqrt{\frac{\sum_{w} (\text{neighbouring pixels} - \text{mean})^2}{w}} \]  

(6)

F. Classification

For the classification of the input images a deep learning method is used that is class of deep neural network known as CNN (convolutional neural network). It is used to analyze the visual imagery and have different applications such as image classification, image and video recognition, object detection, feature and medical image analysis. In the last step the approach of CNN is applied which can categorizes and locate the cancerous nodules. The input image is fed into the network which consists of different layers including input layer, hidden layer and the output layer.

![Fig. 5 Simple diagram of artificial neural network showing three layers input layer, hidden layer, output layer.](image-url)
IV. RESULTS AND DISCUSSION

The proposed approach was implemented in the MATLAB. The common tools to analyze the performance of the system are sensitivity, specificity and accuracy which gives the desired result. These parameters are obtained after the CNN approach is applied for the classification. These parameters can be described on knowing the basic terms false positive, false negative, true positive and true negative which were used to calculate the result. The brief description of the parameters obtained through them is as follows:

Sensitivity is defined as the relationship between the true positive and false negative result obtained whereas true positive (TP) are the actual nodules which are cancerous and are correctly identified by the system. False negative (FN) is when the result obtained is a negative result but the expected result was a positive result. This means nodules which are missed by the system. Thus it can be calculated as given below:

\[ \text{Sensitivity} = \frac{TP}{TP + FN} \]

Specificity is the relationship between the true negatives and false positive where false positive (FP) is when the result obtained is a positive result but the expected result was to be a negative result. This means the non-cancerous nodules are detected as cancerous nodules. True negative (TN) are the nodules which are non-cancerous and are correctly identified as non-cancerous by the system. Thus it can be calculated as given below:

\[ \text{Specificity} = \frac{TN}{TN + FP} \]

False positive rate is the rate obtained after every single scan and can be calculated by the formula given below.

\[ \text{False positive rate} = 1 - \text{Specificity} \]

Accuracy of the system is analyzed by the below given formula that is the sum of true positive and true negative upon the total sum of all the terms that are true positive, true negative, false positive and false negative.

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \]

The proposed algorithm uses the computer vision and neural network tool box for implementation. The interface of lung cancer detection is shown in the Fig. 7. In the figure, all the techniques performed and tumor portion marked image is illustrated. The system will be trained to generate the desired results. The training of the neural network system is shown in the above given Fig. 8.

![Fig. 6 Flowchart for proposed methodology](image)

![Fig. 7 Interface of Lung Cancer Detection](image)

![Fig. 8 System Training](image)

![Fig. 9 A graph showing the best training performance of the neural network for a single image scan.](image)
In the Fig. 9, the graph shows the best training performance of the neural network system for a single image scan. The best performance obtained for the above shown scan is at epoch 3 value. One epoch is defined as one backward pass and one forward pass of the training. This graph shows how to plot training performance of the system. Usually the error is reduced after more epochs of training. Finally the system give the parameters in the terms of sensitivity, specificity, and accuracy which are discussed below.

As shown in the Table. I the sensitivity of each image per scan has shown good performance. The overall sensitivity achieved by the system is 90.39%. The specificity can never be 100% because with the increasing sensitivity the false positive rate per scan will also increase. The performance shown by the CNN approach is good and gives better result than SVM approach. And also the proposed system consumed less time to detect the cancerous nodule on each image with less false positive per scan.

Table. II Specificity analysis

<table>
<thead>
<tr>
<th>Image Number</th>
<th>SVM Approach</th>
<th>CNN Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>83.5</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>84.52</td>
<td>105</td>
</tr>
<tr>
<td>4</td>
<td>85</td>
<td>99</td>
</tr>
<tr>
<td>5</td>
<td>81</td>
<td>97</td>
</tr>
<tr>
<td>6</td>
<td>81</td>
<td>93</td>
</tr>
</tbody>
</table>

As shown in the Table. II the specificity is analyzed for the different images. The overall specificity achieved by system is 96.6% which shows a good performance of the CAD system proposed. The above given graph in Fig. 10, shows the comparison of the accuracy analysis of the existing and the proposed approach. The system is tested on different number of images and it is analyzed that CNN give best results as compared to SVM as a classifier.

Table. III Accuracy analysis

<table>
<thead>
<tr>
<th>Image Number</th>
<th>SVM Approach</th>
<th>CNN Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>82</td>
<td>92</td>
</tr>
<tr>
<td>2</td>
<td>84.5</td>
<td>93.5</td>
</tr>
</tbody>
</table>

The comparison of accuracy achieved for each image per scan is shown in the Table. III. The overall accuracy achieved by the system is 98.08%. The system achieved better results by the CNN approach in the terms of accuracy, sensitivity and specificity. Thus the overall performance of the system was good in all the terms of parameters.

V. CONCLUSION

Although recent advances have been made in the diagnosis and in the treatment of lung cancer but it lacks due to early diagnostic tools. For this CAD systems have been proved to be very helpful in the medical field. CAD is a technology which estimates the features that represents the specific disease process in the medical images and marks the area of concern which needs to be treated. Lung nodules is a growth of abnormal cell tissues which can spread across the body if not treated on time. These growing cells can be malignant (cancerous) or benign (non-cancerous) in nature. These cells can multiply and increase at a very high speed. The early diagnosis is the only remedy to prevent the number of deaths due to this dangerous disease. In this research a CAD system is proposed to detect the lung nodules based on CNN (convolutional neural network) classification and different image processing techniques. The proposed method was performed on CT scan image as input image which gave a clear difference between the true positive and false positive nodules using GLCM feature extraction algorithm. The classification using convolutional neural network identified the cancerous nodules and non-cancerous nodules from the input image. This method improved the overall performance of the CAD system by achieving 98.08% of accuracy, 96.6% of specificity and sensitivity analyzed is about 90.39% with only 0.985 false positives per scan. The time taken by the system on each image is 0.054 seconds. The proposed approach is implemented in MATLAB. Such CAD systems could be very helpful in assisting the radiologists as the approach consumed very less time for the detection of lung nodules with a high accuracy.

REFERENCES


AUTHORS PROFILE

Sayana Sharma is currently pursuing M.tech in Electronics and communication stream as final year student from Punjabi University, Patiala. Her areas of interest include: Image processing and neural networks.

Mandeep Kaur is an Assistant Professor and is currently pursuing her Ph.D in Punjab University, Patiala. Her areas of interest include: Signal processing, embedded systems and Digital Image processing.

Dr. Deepak Saini is an Assistant Professor in Punjabi university, Patiala. His areas of interest include: microelectronics, nano-technology and image processing.