Diagnosis of lung carcinoma using computed tomography images

Sagar V*, Akshatha Prabhu
Department of Computer Science, Amrita Vishwa Vidyapeetham university, Amrita School of Arts and Sciences, Mysuru, Karnataka, India

Article History:
Received on: 13.04.2019
Revised on: 22.07.2019
Accepted on: 26.07.2019

Keywords:
Image Pre-processing, Nodules Feature Extraction, Computed Tomography, Image Segmentation

ABSTRACT
Lung carcinoma disease is a serious leading death every day. Early recognition and treatment can save cancer affected the human's life. Lung carcinoma is the most hazardous and dangerous disease. Computed Tomography scanning images are the most commonly using scanning technology in the health care industry. Computed Tomography scan images are used by doctors to analyze and identify the cancerous lesion present in the lung CT scanned images. The CAD system is a powerful tool for the radiologist to diagnosis the lesion cells in an accurate way. Computer-aided techniques and machine learning techniques use digital image processing methods to implement CAD systems. The prime aim of this investigation and experimentation is to analyze the different CAD techniques, recognizing the best-developed methods and collecting their disadvantages, faults. Analyzing the drawbacks from the previously proposed model, we are implementing new techniques to overcome the drawbacks analyzed with best accuracy results. For developed methods, the feasibility study is carried on each and every step, improvements can be made were listed out. The analysis shows that having a few low accuracies and some have high accuracy but not having 100% accuracy. In this research, K-means algorithm is used to segment the images. K-means algorithm is unsupervised learning, which is used when we have unlabeled dataset, the project aims to increase the accuracy of up to 100%.

INTRODUCTION
Lung carcinoma which commonly uncontrolled growth in lymph nodes of the lung. The common symptoms are like coughing, weight loss, the variation of breath and chest pain. In this paper, the proposed system implemented the watershed segmentation for diagnosis of cancer in the lung CT scan images and SVM is used as a classification tool for classifying as Malignant or benign cancer of detected nodules. In this model, 92% accuracy was produced. But this model did not mention various cancer stages (Makaju et al., 2009). In this model, the pre-processing of the lung of CT scan image is done through clearing noisy data available on the image by using Gaussian filter and then implemented segmentation methods using the different thresholding techniques to produce the good method for the segmentation step. This project more concentrated on identifying lung cancer in the earlier state. In this paper, image improvement method is adapted to detect earlier stage illness identification and accuracy is considered as code factor in this model. Gabor filter within Gaussian rules is used as a pre-processing method. The proposed technique implements region of interest segmentation principles to produce efficient and good results compared with other techniques. Mask-labelling is done for
more accuracy and robust operation (Al-Tarawneh, 2012). In this paper they implemented a classification technique using the Deep Fully Convolutional Neural Network (DFCNet). DFCNet is a general classifier which is used for detection and classification of digital image processing technique images. But, in this paper DFCNet is used to automatic diagnosis and classifying of the pulmonary lesion nodules in the CT images. In the initial process, Malignant or Benign cancer and normal lung image classification were done (Masood et al., 2017).

Related Works

The author Zia et al. (2018), This paper proposed censorious analysing of present nodules detection method of lung cancer detection. Gathering the information on the present trends and identifying the subsequent challenges. However, upcoming research is able to improve existing methods or implement new methods that provide a conclusion with greater sensitivity and lower FP rate, and ultimately provide simple combination with PACS and EMR systems. The author Kashyap et al. (2016), In this paper, has mainly concentrated on the digital image processing techniques to find or identify cancerous cells or lesion in the Computed Tomography image. The pre-processing method for a lung image performed by using a median filter because to clear unwanted noisy data presented in the available images and then in segmentation process morphological operators is implemented and clearly drawn the affected tumour region. Later cognitive approaches are used to find the tumours then feature extraction step is carried out to calculate the features. The author Skourt et al. (2018), In this paper, has explained about a lung cellular tissue segmentation method by implementing U-net architecture and has generated the faultless segmentation process with the value of 0.9502 Dice-coefficient indexes. Favourableness of this proposed paper is uniformly this method are developed and also useful to the wide range of various health care image segmentation process can be implemented. The objective in this paper is U-net architecture segmentation is proposed that to segment lung nodule, provides the best accuracy to their project and can be adapted to the other project also. The author Nisha and Maheshwari (2015), In this paper, the writer implemented the lung cancer detection technology to find early-stage identification of lung cancer through calculating the number of steps. The method implements by extracting the lung region on lung CT image using various image processing techniques including binary image, erosion and Gaussian filter. Once the extraction step is completed, the region growing segmentation technique is implemented on extracted lung regions. Then feature extraction is done by calculating the shape, area, and perimeter on an extracted lung region. Later using the values of features extracted, the lung image is classified as lesion cancer or non-cancerous by an artificial neural network using backpropagation is implemented. The main advantages of the artificial neural network calculate the required output otherwise it backpropagates and adjusts the weights and it process again until required output and actual outputs are similar. This paper implements the best ANN technique used to the classification of lung cancer on CT scan images. This model helps the physician and radiologist to identify the nodules that increased the sensitivity in the diagnosis. The author Miah and Yousuf (2015), In this paper, the proposed method initially converts the grayscale images into binary thresholding. After segmentation process is done then feature extraction method is implemented, and then using those features extracted, training of the neural network is implemented and testing the neural network. The proposed technique accurately detects lung cancer from a CT scan image. The advantage of this model is expected output is reached. This model test 150 kinds of lung CT scan images and prevailed the result in which the success rate was 96.67%. This proposed method can be used in the detection of breast cancer, brain tumour etc. The author Image and Nakajima (2018), In this paper latest and creative bronchoscopy techniques is adopted for recognition of early stage lung cancer: The bronchoscopy optical biopsy might possibility restore conventional biopsy for tissue detection. Navigational bronchoscopy developed is an advantage step in early stage lung cancer recognition. Classification works as powerful to the diagnosis of malignant and benign cancer. It reduced the big procedure to detect the cancer lesion. BMC and EBC play helpfulness for early recognition of lung lesions. The author Kuruvilla and Gunavathi (2014), In this paper, they implemented a CAD technique and segmentation is done to an entire dataset images and from segmented images parameters are calculated. The standard parameters are GLCM features is calculated and then used to specify as benign or malignant. For the classification method, a combination of two methods are used to classify, they are Feedforward neural networks and feed forward back propagation neural networks are used. Feedforward back propagation network gives more accuracy. The results of classification accuracy were 91.1%. The author Arulmurugan et al. (2018), In this paper for the segmentation method they have used region of interest which calculates from slices of the lung images. Wavelet
feature is extracted from calculating from the GLCM method. The four training functions were used to construct the neural network (NN). The accuracy is increased by using the traindx training function. From the implemented NN feed forward back propagation classification method helped to generate Accuracy of 92.61%, specificity calculated is 100% and the result of sensitivity obtained is 91.2% then 0.978 is the identified mean square error. This proposed technique helped to diagnosis the early stage of lung cancer. The author Wang et al. (2018), In this paper they have implemented a pulmonary nodules CAD algorithm using a semi-supervised ELM (SS-ELM) technique, which contains class features of both labeled and unlabeled features. Then features are inputted to training and calculated as benign or malignant cancer. First, the Haralick feature model is used to identify lesion nodules locations in the lung images. Dataset of lung images is downloaded from LIDC-IDRI in TCIA website which consists of 1018 set of CT scan images. The results are calculated by using SVM and SS-ELM obtained more accurate performance. The author Rendon-Gonzalez and Ponomaryov (2016), In this paper, model is developed by implementing Region of Interest (ROI) method to segment the CT scan images using the prior information and HousefieldUnit. The classification method used in this paper is the support vector machine (SVM). To classify as benign or malignant, shape features and textural features are calculated. Textural features are extracted to train the model to classification method and classified using a support vector machine to recognize cancer stage. The author Sangamithraa and Govindaraju (2016) has proposed that segmentation method is developed through K mean unsupervised learning technique. Pre-processing step is done with K mean unsupervised learning technique. Pre-processing step is done with median filtering method, which is applied to debug noisy data available in images that improves the accuracy. In K-means segmentation algorithm on the basis of similar characteristics, it forms clusters in an image. And classification is done by implementing the back propagation NN method. For classification purpose shape features are calculated using equations of area and perimeter were calculated. Features are calculated by using the GLCM method. This model proposed an increased accuracy in the diagnosis.

MATERIALS AND METHODS

According to the analysis of previously developed methods, analysis shown that accuracy is poor and time requires more than the estimated time to identify and classify. Therefore, the new methodology is implemented to overcome the drawbacks and limitations were there in the previously developed techniques and workflow of the project is shown in Figure 1 and implemented method is shown in Figure 3. The prime robustness features of this proposed method are listed below,

1. Increased the exactness of cancer diagnosis level more than the current proposed and implemented models.
2. Malignant and benign images are recognized accurately.
3. Image enhancement was done in an appropriate way by applying histogram equalization, so pre-processing technique obtained with good result and error causing data are removed in the pre-processing step

Histogram equalization technique is developed in a pre-processing step to improve pixel quality. From the pre-processed image, then the processed image is K-means algorithm is implemented and Otsu thresholding method is developed in the segmentation phase. Then cancer nodule is identified. Texture features is calculated by GLCM method which gives more accuracy method in feature extraction steps from a segmented image. However, then using features calculated by GLCM method and trained data which is imported are compared and results are shown as either benign cancer nor malignant cancer in classification step by Implementing SVM technique.

Image acquisition

Image acquisition is a process in the Digital image processing can be briefly explained as the form of receiving the image from various sources, commonly a hardware-based source, so it can be sent through any processes required to occur next process. Implementing image acquisition in the digital image processing is the first step in the implementation and without an image uploading, no processing is possible for the next process.

Image Pre-processing

In this model pre-processing is developed by histogram equalization technique on lesion images. Smoothing and filtering of an image is done in the pre-processing method to remove noisy data and sharpening the region of interested elements. Noises may cause false detection of cancer nodules. Therefore, noises have been filtered for the accurate detection of cancer in the images.

The histogram equalization method commonly increases the global contrast of grayscale images. From global contrast adjustment, the intensity values are altered and implemented the enhanced
images. The low level brightness is converted or implement into higher brightness level. Histogram equalization implements more brightness images so images are perfectly processed the intensity pixel values.

The technique is more useful in images with bright or dark in both backgrounds and both foregrounds.

**Segmentation**

**K-means segmentation**

K-means segmentation is the unsupervised technique model which partitions or divide a gathering of data using K value which is number initialized during execution of the program. In K-means segmentation, Euclidian equation is used to calculate the distance between the object present to form or to group the similar data and forming the clustered element. Figure 2 represents K-means Algorithm.

In the proposed method, we have initialized k value as 3, so it forms 3 different similar objects and clustered into the single image

**Otsu segmentation**

Otsu segmentation coverts the grayscale image into a binary image. This segmentation uses the intensity values to segment the images. If the intensity value is above 128 intensity value then it converts to the white pixel value, if the intensity value lies below 128 intensity value then it converts to the black pixel value.

Euler number(EN) is calculated after the Otsu segmentation method is done

\[
\text{Euler Number} = \text{the number of objects present minus integer values of hollow in an image.}
\]

Euler Number of an input image is 111 value.

**Features extraction**

Features extraction are used to distinguish applicable information for solving the computational problems in an application. There are several different types of features. In this new proposed methodology, features are discovered by Gray Level Co-occurrence Matrix (GLCM) feature extraction method using formula represented in Table 1.

Table 2 shows the results of Gray Level Co-occurrence Matrix (GLCM) feature extracted value of Figure 8 of k-means segmented image.

In this step, 13 texture features are calculated by GLCM method like Inverse Difference Moment intensity and so on. These features are stored in separate
file and helped as training features to develop classification.

After K-means segmentation method, GLCM method is implemented to extract the Gray level feature are extraction and stored in a dot(.) mat file which is further used for classification technique to produce final output.

Table 3 shows the classification results of few images after all the implementation is done.

**Classification**

In this phase identifying and displaying either malignant nor benign stage of the detected nodules. Support vector machine (SVM) method is developed to classification. SVM is a superintend instrument learning and training technique. SVM classification uses Training dataset to classify the images. The training dataset consists of features of the entire dataset and stored in a dot mat file. Comparing the features extracted from the GLCM method and training dataset, SVM classifier classifies the images as benign or malignant

**Implementation**

**Dataset**

The Cancer Imaging Archive (TCIA) is an administration which de-distinguishes and has a huge document of therapeutic pictures of disease open for public download. In TCIA there are 7 varieties of dataset among them LCTSC is the one we have chosen.

Lung CT segmentation Challenge 2017 (LCTSC) has cancer affected lung CT images. In Lung CT segmentation Challenge 2017, 4.8GB dataset is available. I have download 4.8GB complete dataset. It has more than 8000 images. The dataset consist of 512*512 pixel images and all images are in DICOM format. From DICOM format images, the dataset is converted into JPG format by using the tool MicroDicom viewer software. MicroDicom software tool converted all the images in the dataset without causing or missing of a single image and gives more accuracy for converting images to JPG format.

**RESULTS AND DISCUSSION**

Figure 4 represents the original image of CT scan images.

Figure 5 represents Wiener filter applied to the original image.

Figure 6 represents Histogram equalization preprocessing method applied to the original grayscale image.
Table 1: Feature extraction formula

<table>
<thead>
<tr>
<th>Features</th>
<th>Formulas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$M = \mu = \frac{1}{XY} \sum_{g=1}^{x} \sum_{h=1}^{y} p(g, h)$</td>
</tr>
<tr>
<td>Variance</td>
<td>$V = \sigma^2 = \frac{1}{XY} \sum_{g=1}^{x} \sum_{h=1}^{y} [p(g, h) - \mu]^2$</td>
</tr>
<tr>
<td>Energy</td>
<td>$E = \sum_{h=1}^{y} [p(g, h)]^2$</td>
</tr>
<tr>
<td>Entropy</td>
<td>$F = \sum_{g} \sum_{h} p(g, h) \log (p(g, h))$</td>
</tr>
<tr>
<td>Inverse Difference Moment (IDM)</td>
<td>$I = \sum_{g,h} \frac{p(g,h)}{g-h}$</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>$K = \frac{1}{XY} \sum_{g=1}^{x} \sum_{h=1}^{y} \left</td>
</tr>
<tr>
<td>Skewness</td>
<td>$S = \frac{1}{XY} \sum_{g=1}^{x} \sum_{h=1}^{y} \left</td>
</tr>
<tr>
<td>Contrast</td>
<td>$C = \sum_{g=1}^{x} (g-h)^2 p(g, h)$</td>
</tr>
<tr>
<td>Smoothness</td>
<td>$A = 1 - \frac{1}{(g-\mu_i)(h-\mu_j)p(g,h)}$</td>
</tr>
<tr>
<td>Correlation</td>
<td>$B = \sum_{i,j} \frac{(g-\mu_i)(h-\mu_j)p(g,h)}{\sigma_i \sigma_j}$</td>
</tr>
</tbody>
</table>

Figure 3: Proposed model
Figure 7 represents the Otsu segmentation, segmentation method applied on Histogram equalization method to generate Otsu segmentation.

Figure 8 represents K-means segmentation, segmentation method applied on Histogram equalization method to obtain K-means segmentation.

Figure 9 represents Canny edge detection, obtained result on applying to K-means segmented image.

Figure 10 represents Binary gradient mask, obtained result on applying to K-means segmented image.

Figures from Figure 11 to Figure 20 are Dataset.

Total number of lung images (TN) = 10

Number of images with true classification (TC) = 9

<table>
<thead>
<tr>
<th>Feature extraction</th>
<th>GLCM values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast</td>
<td>0.5254</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.9558</td>
</tr>
<tr>
<td>Energy</td>
<td>0.1404</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>0.8615</td>
</tr>
<tr>
<td>Mean</td>
<td>0.9475</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.8850</td>
</tr>
<tr>
<td>Entropy</td>
<td>6.3053</td>
</tr>
<tr>
<td>RMS</td>
<td>1.2603</td>
</tr>
<tr>
<td>Variance</td>
<td>0.7004</td>
</tr>
<tr>
<td>Smoothness</td>
<td>1.0000</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>6.2611</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.7122</td>
</tr>
<tr>
<td>IDM</td>
<td>563.6506</td>
</tr>
<tr>
<td>Serial Number</td>
<td>Images</td>
</tr>
<tr>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>Image 82</td>
<td>Figure 11</td>
</tr>
<tr>
<td>Image 88</td>
<td>Figure 12</td>
</tr>
<tr>
<td>Image 97</td>
<td>Figure 13</td>
</tr>
<tr>
<td>Image 11</td>
<td>Figure 14</td>
</tr>
<tr>
<td>Image 16</td>
<td>Figure 15</td>
</tr>
<tr>
<td>Image 54</td>
<td>Figure 16</td>
</tr>
<tr>
<td>Image 123</td>
<td>Figure 17</td>
</tr>
<tr>
<td>Image 129</td>
<td>Figure 18</td>
</tr>
<tr>
<td>Image 108</td>
<td>Figure 19</td>
</tr>
<tr>
<td>Image 97</td>
<td>Figure 20</td>
</tr>
</tbody>
</table>

Figure 11: Begin

Figure 12: Begin

Figure 13: Begin

Figure 14: Begin
Figure 15: Begin

Figure 16: Malignant

Figure 17: Malignant

Figure 18: Malignant

Figure 19: Malignant

Figure 20: Malignant
Number of images with false classification (FC) = 1

Accuracy = \( \frac{TC}{TN} \times 100 \)

\( \text{Accuracy} = \frac{9}{10} \times 100 = 90\% \)

Therefore, accuracy is increased to 90% in this research paper.

Along with strengths, this model has some few weaknesses too. They are listed below:

1. In this model, accuracy is increased from 86% to 90% in an implementation phase but not reached to 100% accuracy.

2. SVM classifier didn’t identify nor classifies as various types like stage 1, stage 2, stage 3, stage 3A, stage 3B and stage 4.

**CONCLUSIONS**

According to the current developed models, there is no satisfactory accuracy obtained in the proposed methodologies. Hence, the new proposed model is implemented. The implemented model can be applied to identify the lesion cells in nodules of a lungs pictures by K-means segmentation to the recognition. SVM technique is applied to find and classify the cancer images as two types of lung cancer, either benign or malignant cancer. The new proposed methodology detects cancer with more accuracy than compared to current developed models and SVM classification method accuracy has increased from 86.6% to 90%. In general, we have developed an improvement in the implemented system when we compared with the currently proposed models. Although, in this proposed model limitations exits because we didn’t find nor organizing the various cancer phases as stage I, stage II, stage III and stage IV cancer. Therefore, in future enhancement different cancer stages like stage I, stage II, stage III and stage IV cancer can be implemented and also using different pre-processing method accuracy can be increased by removing false objects present in the CT scan images. The main purpose of using K-means algorithm because when research is, K-means segmented images leads to giving the little bit of accuracy less and identified, why accuracy is giving less and implemented by giving accurate K values and correct pre-processing methods of images are done so the k-means segmentation gives the accurate segmented values. If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. K-means produce tighter clusters than other hierarchical clustering algorithm. GLCM feature extraction method is used mainly because K-means produces different clusters in the same image from that segmented images, we have applied the GLCM method to increase the accuracy. GLCM method provides 13 different features from the segmented images in which it helps us to increase accuracy, which identifies the texture features from the images. SVM method is used for classification method because in previously implemented projects in lung cancer classification are producing less accuracy after the image segmentation method after research got to know that using SVM after K-means segmentation method gives more accuracy than the current best model. After implementing this project, the accuracy is increased from 86% to 90% in the lung carcinoma detection and classification.

**ACKNOWLEDGEMENT**

I am very grateful to my parents, teachers, instructors and friends whom helped me with the courage and enthusiastic to write this article. This research project is supported by our college in Mysore, Karnataka, India.

**REFERENCES**


