Artificial neural network for lung cancer detection using CT images

Shankara C
Department of Electronics Engineering, Faculty of Engineering and Technology, Jain University, Bangalore, Karnataka, India
Email: shankara151987@gmail.com

Hariprasad S A
Department of Electronics Engineering, Faculty of Engineering and Technology, Jain University, Bangalore, Karnataka, India
Email: sa.hariprasad@jainunversity.ac.in

Abstract---Lung Cancer is the second most serious disease in today’s world due to which the mortality rate is increasing every year. Accurate and early identification of cancer in the lung could help people live longer. Medical image processing has a significant impact on the recognition of lung tumors using Computer Tomography (CT) scan images. CT scan images provide the complete imaging of tumor development inside the lungs and are extensively used. A lung cancer diagnosis is usually done manually by skilled specialists, and while these approaches are particularly useful in advanced stage detection, it is also a time-consuming operation that is highly reliant on the person. This increases the risk of detecting inaccuracy due to human error procedures, necessitating the use of an automated system. As a result, it is necessary to identify lung cancer using an automated approach to decrease human manual error and enhance the accuracy and convenience of the process. In this proposed method image processing techniques and an artificial neural network (ANN) were employed to create an automatic method for accurate lung cancer detection. In the first step, image noise reduction, the histogram for image enhancement, Thresholding for image segmentation, object edge detection, morphological operations (dilation), and Gray Level Co-Occurrence Matrix (GLCM) for feature extraction are used. The ANN is built in the second step using a machine learning technique. The proposed system’s performance was analyzed and an accuracy of 94.89% was achieved.

Keywords---artificial neural network, classification, histogram equalization, segmentation, thresholding.
Introduction

Lung cancer is the most serious and pervasive disease in the world today, with a rapidly rising fatality rate. The mortality rate from lung cancer is expected to reach 2.45 million people by 2030 (that is 39 percent higher than the number of cases recorded in 2018) (Abdillah et al., 2017). The count of people diagnosed with lung cancer is expected to rise to 32, 99,640 by 2040. Lung cancer claims the lives of around 1.3 million people worldwide each year, with an assessed 1.8 million lung cancer new cases. The overall number of cancer cases in the lungs, recorded around the world in 2018 was around 2.09 million (which includes 1.36 million patients who were males, as well as 0.73 million cases, were in women) (Chaturvedi et al., 2021). Also stated that the fatality rate from lung cancer was 1.76 million in 2018 (Jony et al., 2019). Lung cancer was responsible for roughly 19.4 percent of all cancer-related deaths in 2018 (Kaucha, Deep Prakash, P.W.C. Prasad, Abeer Alsadoon, 2017). Lung cancer affects one out of every 68 males and one out of every 201 females in India, according to estimates. Lung cancer affects one out of every 101 people, according to statistics. In the year 2020, 71,788 new cases occurred in men and 26,490 new cases occurred in women (a total of 71,788 new cases). The total cases recorded in India in 2018 were 67,795 (which includes 48,698 males and 19,097 females). There were approximately 45,363 recorded deaths (comprising 27251 men and 18,112 women). Lung cancer patients (at all stages) have a survival rate of about 14% and the time span of survival is for about 5–6 years only (S. Kalaivani, Pramit Chatterjee, 2017). Smokers have a high possibility of lung cancer hence lung cancer is one of the first types of cancer leading to maximum deaths in both males and females, probably as a result of passive smoking or chewing tobacco. American cancer society states smoking and passive smoking are considered to be responsible for 80% of lung cancer fatalities. The increased rate of pollution in the environment is also the major cause of lung carcinoma (Manika Jha, Richa Gupta, 2020). 85% of lung cancer cases are the result of Non-small cell lung cancer and the remaining 10–15 percent of lung cancer cases are the result of small cell lung cancer (Abdullah et al., 2021).

Doctors can use a computer-aided diagnosis (CAD) system to discover and diagnose anomalies early and more quickly. Lung cancer is a disease that can affect anyone and it can be diagnosed by a variety of methods, including chest radiography, the computed tomography (CT), magnetic resonance imaging (MRI), and cytology (Nadkarni & Borkar, 2019a). However, the majorities of these methods are more expensive, more time-consuming, and have less capability for identifying lung cancer. The majority of these methods for detecting lung cancer in its advanced stages result in poor patient outcomes.

Many studies have presented a variety of strategies for image processing and automatic classification systems, with variable outcomes. In the traditional method of lung cancer detection only image processing methods are used and also more stages in the detection process increase the computational time and error rate (Firdaus et al., 2020). It leads to less classification accuracy, less recognition rate, and a high classification error rate. Hence a new strategy for predicting lung cancer early and more accurately is required. The Neural Network is used to create an integrated framework for predicting lung cancer. Neural
networks deal with a variety of problem-solving strategies in which neurons are trained and tested using a database. The lung cancer traits are retrieved to estimate the cancer stage based on the system's features. Feature selection reduces the number of cancer cells supplied to the computation technique by identifying predicted subsets of cancer cells within a database.

By removing some features, you can improve performance (Begum & Ismail, 2021). Hence to predict lung cancer with more accuracy in less computational time image processing techniques and ANN method are used. ANN is an end-to-end process that takes less computational time and is employed for training and testing of the CT images for lung nodule classification (Chauhan & Singh, 2019). There is very much important to identify cancer by using image processing method and ANN method. In the remaining part, the structure of this document is as follows: The section 2 explain the materials and methods applied in this analysis. The Section 3 explain results and discussions. Section 4 describes the conclusion of this research work.

Materials and Methods

The different researcher applied different method for lung cancer detections. The researcher (Cruz et al., 2016) presented a neural network with back propagation and genetic algorithm method for lung cancer detection consists of different stages. In this method image pre-processing, the feature extraction, and the classification are used. The images downloaded from LIDC database are filtered to eliminate the noise using a grayscale conversion algorithm. From the filtered images, features are extracted using a genetic algorithm. An ANN has to be trained with a training dataset for learning and performing classification. The back-propagation neural network method was applied for classifying the images as cancer or non-cancer using extracted features. The main limitation of this system is the hybridization of genetic algorithm fail to predict lung cancer accurately and also good accuracy is not achieved.

The author (Ziyad et al., 2019) presented a method that uses the LIDC dataset for lung cancer detection utilizing chest radiographs and feed-forward artificial neural networks for categorization. The method uses histogram equalization in the preprocessing stage to remove the undesired information or noises in the CT images. After pre-processing the image set is applied with segmentation methods like intensity-based and discontinuity-based methods to segment the images and it will give only the necessary information from the images. For segmented images, feature extraction is performed, and features like area, diameter, irregular index, mean, median, and others are extracted. For classification, the artificial neural network is applied which divides the images into cancerous and non-cancerous. The system produced an accuracy of 92% for the identification of lung tumors and the limitation of this system is it is unstable when the size and position of the lung image are changed.

The researcher (Pandiangan et al., 2019) proposed artificial neural networks for lung cancer detection which consists of five stages: noise reduction, enhancement of an image, lung image segmentation, objects edge detection, and the tumor boundary recognition. In the noise reduction stage, the system uses the grayscale
method for the removal of noises, and the spatial domain filtering technique with averaging mean is applied in the image enhancement stage. Afterimage enhancement, the segmentation method is applied to segment the images. By applying the edge detection technique the system will find the edges of all the required points in the input image and finally, the artificial neural network is applied as the image classifier to divide the images into cancer and non-cancer.

The author (Anifah et al., 2018) presented a system that uses GLCM for feature extraction and classified the CT images into normal and cancerous images using a back propagation artificial neural network. In the proposed system the images or the data are collected from the cancer imaging archive database. For the image pre-processing they used median filter and adaptive histogram equalization to eliminate the unwanted noises from an image and to improve the intensity of the image. Then the region-based techniques are used in the segmentation stage to collect the essential information and other unnecessary data will be removed. The features such as the contrast, the correlation, and the variance are gathered from the given segmented image by applying the GLCM method. Finally, the ANN along with the back-propagation is used as the classifier and the system achieved an accuracy of 80%. The system can implement various combinations of filtering and segmentation methods to increase accuracy.

The researcher (Nadkarni & Borkar, 2019a) proposed a lung tumor detection system for CT images based on image processing that uses a dataset from the cancer imaging archive (TCIA). In the pre-processing step, the median filter is employed to eliminate unwanted noise in the lung image. Then mathematical morphology method is used for the segmentation of the images. By using this method the required regions are gathered from the input image. Three geometrical features like area, eccentricity, and perimeter and four textural features such as energy, contrast, homogeneity as well as correlation are extracted in the feature extraction stage. Finally the SVM classifier is applied to divide the images into normal and abnormal lung images.

The author (Kamble et al., 2016) uses lung CT images as input dataset to the Computer Aided Diagnosis (CAD) and stationed on an algorithm it helps the analyst to perform an image inquiry. The median filter is used to eliminate noise in the input images. With the aid of neural networks and Self-Organizing Maps (SOM), the created CAD system favours segmentation of nodules on the lobes and ensures classification between benign and malignant nodules. Binarization and GLCM methods are used for feature extraction. Improved method of segmentation planted on soft computing will be used for lung cancer detection.

The researcher (Jitender, n.d.) developed a system that includes image acquisition, the pre-processing, thresholding, feature extraction, binarization and neural network detection. For experimentation about 150 CT images are considered which consists of both male and female lung images with 512 x 512 pixels in size. Two-dimensional OTSU method and two morphological operations like dilation as well as erosion are applied to compass individual lung images and also to eliminate unwanted parts using segmentation. Using GLCM method features like Energy, contrast and Entropy are extracted. For classification purposes, the
artificial neural network is performed very well to detect cancer in the lung CT images.

The author (Kaur et al., 2021) applied SVM to train the dataset. In this method, image processing methods are applied to remove noise and other limitations used a classifier to extract the various features of images from the labeled database, used machine learning to train 80% of the sample database. Later test the network with the remaining 20% of samples, affix the trained network to on the scanned images and verify the results, calculated the efficiency of the method in comparison to other methods and finally visualize the results. In the future work the efficiency of the current algorithm can be evaluated with many other better algorithms for lung cancer detection. Also various other known objective parameters like eccentricity, entropy can be used for checking the implementation of the current algorithm.

The researcher (S. Kalaivani, Pramit Chatterjee, 2017) proposed an ANN and image processing methods were applied to create self-operating process for detecting lung cancer. They used data from The Cancer Imaging Archive’s Lung CT-Diagnosis database. There are 4682 similar photos in the database. Gray scale conversion, histogram equalization, the thresholding as well as feature extraction are the methods involved in image processing The BPN is made up of six input neurons, two output neurons, and twelve hidden layer neurons that were trained using 70 images. The result reveal whether tumor in the lung is cancerous or benign. It was discovered through a high pass filter. They used salt as well as pepper noise to decrease the amount of distortions in the image, resulting in reduced distortion. Finally, the image is pre-processed by sending salt as well as pepper noise through a median filter, it removes noise completely while restoring edges. P-tile thresholding is used to detect cancer cells. In this procedure, the watershed algorithm will identify the parts of the infected lung afflicted by cancer, allowing us to distinguish cancerous lung tissue from normal lung tissue.

**Proposed Methodology**

Many image processing algorithms can be utilized in the medicinal industry to detect lung disease. There are four primary stages involved in improving the identification of lung tumors in CT images. Various methods are
Figure 1. Proposed methodology cancer detection in lung

used at each step, resulting in varying accuracies in lung cancer detection. In our proposed method first, the lung CT images are pre-processed to eliminate any background noise. The image is then segmented to obtain a Region of Interest (ROI). Finally, features such as energy, entropy, homogeneity, and variance are extracted using feature extraction algorithms. Finally, various classification algorithms are applied to the retrieved elements of the lung CT image to determine if it is malignant or not as shown in Figure 1.

Dataset

The image collection is the first stage in medical image processing. This proposed automated process system took a series of CT scan images as input, which were analyzed to categorize lung nodules. The LIDC Database (Lung Image Database Consortium Computing) (Wu et al., 2020) is a service that de-identifies and archives a vast collection of cancer-related lung images. A total of 3748 lung CT images are used in the experimentation, which consists of 2923 cancerous lung CT images and 825 normal lung CT images. For training 3611 lung CT images are used and for testing 137 lung CT images are used.

Image Preprocessing

Because in image processing, intensity images are used as input and output, pre-processing is the frequent term for actions involving images at the lowest level of abstraction. Image processing is used to improve visual data by removing distortions and enhancing certain image elements that are relevant for further processing. It is necessary to preprocess images to improve contrast, sharpness, and separation of background noise. As a result, several approaches such as smoothing and augmentation are used to achieve the desired image. The histogram equalization (Kumar, 2020) approach is used to improve the brightness of an image. It is the process of smoothing out an image’s histogram. It reduces the amount of noise and other minor changes in the image. When compared to the original image, histogram equalization results in a sharper and
crisper image. Sharper borders and highlighted items in the image are examples of this. It produces an image with a higher dynamic range and contrast. As a result, we used histogram equalization to the grayscale image to boost its contrast.

**Image Segmentation**

Segmentation is the technique for segmenting image data into distinct sections with similar features. Thresholding is employed for segmentation in the proposed method. Thresholding is a method of segmenting (Evangeline, 2020) an image that involves eliminating data from pixels in a grayscale image that are above or below a specified level. It keeps the necessary information in the image while discarding the unnecessary data.

**Feature Extraction**

In image processing, feature extraction technique is the most used method for extracting features. The most crucial information in an image is conveyed by its edges and lines, which must be retrieved. The CT scan images contain all information of the lung regions, with feature extraction only the required information is extracted to classify either cancerous or non-cancerous. It describes the qualities that remain in the binary image. The neural network is trained and, finally, decisions are made using the range of values for these qualities. Using the Gray Level Co-occurrence Matrix (Mall et al., 2019), features like energy, contrast as well as entropy are retrieved from the resulting image and offered for classification.

**Image Classification**

Neural networks are one of the learning algorithms used in machine learning. They are made up of various layers for data analysis and learning. Each time the network processes input, it learns and assigns weights to the connections between the individual neurons. They are made up of various layers for data analysis and learning. Neural Networks are a form of a multiprocessor computer system with high-level interconnects, simple processing elements, adaptive interactions between elements and messages. These neurons work together in a distributed manner: (i) to learn from input information, (ii) to coordinate internal processes, and (iii) to optimize the outcomes. A neural network's goal is to teach it to recognize patterns in data. For classification, the proposed method employs a back propagation artificial neural network (Nguyen et al., 2020) as shown in Figure 2.

These are data processing systems with more neurons for highly interconnected processing. Three layers make up a network, among the layers are the input layer, hidden layer, as well as output layer. For a given layer of neurons, the error is the difference in the outputs and also their corresponding inputs in a back propagation network. At all layers of the network, a sigmoid training function is used. The mean squared error (MSE) is utilized to evaluate the network's performance. After each epoch, the network itself is trained by adjusting its weights in response to the error. In this method, the network has ceased to
update the weights as soon as it achieves the lowest MSE.

![Figure 2. Artificial Neural Network](image)

**Results and Discussions**

**Image Preprocessing**

The input images are collected from LIDC–IDRI dataset are in DICOM image format, which are converted into JPEG format. The input images are first transformed from RGB to grayscale images. The histogram equalization approach is used to boost the contrast of grayscale images. Later binarization method is applied which converts the grayscale image into a black or white image. All the output images obtained after preprocessing are shown in table 1.

<table>
<thead>
<tr>
<th>Image Id</th>
<th>Original Image</th>
<th>Enhanced Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imag-001</td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
</tr>
<tr>
<td>Imag-002</td>
<td><img src="image" alt="Image" /></td>
<td><img src="image" alt="Image" /></td>
</tr>
</tbody>
</table>

**Image Segmentation**

The preprocessed image is segmented using thresholding to get the required region of the lung. For segmented image morphological operations like filtering, dilation, and image filling are applied. The output images obtained after
segmentation operation, filtering operation, and image filling operation are represented in table 2.

<table>
<thead>
<tr>
<th>Image Id</th>
<th>Enhanced Image</th>
<th>Segmented Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imag-001</td>
<td><img src="image1" alt="Enhanced Image" /></td>
<td><img src="image2" alt="Segmented Image" /></td>
</tr>
<tr>
<td>Imag-002</td>
<td><img src="image3" alt="Enhanced Image" /></td>
<td><img src="image4" alt="Segmented Image" /></td>
</tr>
</tbody>
</table>

**Table 2**
Output results of segmentation

**Feature Extraction**

The GLCM method is utilized to extract attributes from the generated image such as energy, contrast, and entropy. The features extracted from the GLCM method for tumor detection are shown in the table 3.

<table>
<thead>
<tr>
<th>Features Extracted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Image id</strong></td>
</tr>
<tr>
<td>Imag-001</td>
</tr>
<tr>
<td>Imag-002</td>
</tr>
</tbody>
</table>

**Lung Cancer Detection**

From the given feature set the lung CT images are classified as cancerous images using three-layer artificial neural networks. The result of the affected lung with cancer for image1 is represented in table 4. In the proposed method accuracy is used to assess the system’s performance.

| Classification results for imag -001 and imag -002 |
|---------------------------|-----------------------------|
| **Image Id** | **Original Image** | **Normal Image/Cancerous Image** |
| Imag-001 | ![Original Image](image5) | ![Normal Image/Cancerous Image](image6) |
| Imag-002 | ![Original Image](image7) | ![Normal Image/Cancerous Image](image8) |
In the proposed methodology first, the input lung CT images are grayscale images that have been transformed from RGB images. For images in grayscale to obtain a smoothed image, histogram equalization is used. The pre-processed images are segmented using the thresholding method to get the region of interest. For segmented images, filtering is used to eliminate noise in the image. For filtered image morphological operations like dilation and the image, filling operation is applied. From the resultant image, different features like energy, entropy, the contrast, and homogeneity are extracted using the GLCM method. Using these features back propagation artificial classifies the lung CT images into cancerous and normal images. The GUI design for cancer detection of normal lung images is shown in figure 3.

Back propagation ANN method is consists of three layers: input, the hidden layer as well as output layer. This network uses 3 iterations and when reaches minimum gradient it classifies input lung CT images into cancerous and non-cancerous image. The neural network training using nntraintool from Matlab is represented in figure 4.

Figure 3. GUI design for cancer detection of normal image
Figure 4. Neural network training of normal image

Figure 5. The custom neural network of normal lung image

Figure 6. Performance graph of Normal lung image

The custom neural network for non-cancerous lung image is shown in figure 5. The proposed neural network's performance when applied to non-cancerous lung image is shown in figure 6. The neural network training state graph for non-cancerous lung image is graphically shown in figure 7.
The graph in figure 8 represents the regression graph for the proposed neural network when applied on the normal image in which the output value coincides with the target value where the target graph and the output are almost the same. After the target value is almost equal to the output value of the ANN process, then the sample testing process is done. These test results indicate that ANN has been
successfully trained to recognize samples. The sample has been identified with a known shape value as lung nodule.

In the proposed method accuracy is calculated by plotting the Confusion matrix and getting the values for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Generally for good accuracy True Positive and True Negative should be high when compared to FP and FN values which increase accuracy. The confusion matrix for our proposed system is represented in table 5.

<table>
<thead>
<tr>
<th>True Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>3</td>
</tr>
<tr>
<td>False Negative</td>
<td>True Negative</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

In our proposed method accuracy is used to assess the system’s performance. The accuracy is given by

\[
\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100
\]

Accuracy = 94.89%

<table>
<thead>
<tr>
<th>Author</th>
<th>Method applied</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>S.Kalaivani (S.Kalaivani, Pramit Chatterjee, 2017)</td>
<td>RGB to Grayscale, Histogram equalization, Binarization, Thresholding, Region props function, Back propagation ANN</td>
<td>Accuracy=78 %</td>
</tr>
<tr>
<td>Mahmudul Islam (Islam et al., 2019)</td>
<td>Median filter, Gabor filter Thresholding, Otsu method, SVM, GLCM</td>
<td>Accuracy=78.95</td>
</tr>
<tr>
<td>Prenitha Lobo (Demir et al., 2018)</td>
<td>Adaptive Histogram Equalization with Contrast Limits, Fuzzy c means, Support vector machine</td>
<td>Accuracy=79.166%</td>
</tr>
</tbody>
</table>
From the experimental results, it is decided that method proposed effectively detects lung cancer in CT images compare to other existing algorithms. The comparison of different lung cancer detection methods proposed by the various researchers with the proposed method is shown in the figure. 9.

**Conclusion**

We examined a lung cancer detection system employing different approaches for pre-processing, lung segmentation, and classification of lung CT scan images to improve lung cancer detection performance. We chose CT scan images for this research because they are more ideal for accurate results, and they are the most successful way of lung nodule detection due to their capacity to provide three-dimensional (3D) images of the chest, resulting in increased nodule resolution. In MATLAB, we created a lung cancer detection system utilizing ANN. An input layer, the hidden layer as well as output layer makes up the network's three layers. Our model performed well enough to correctly classify the given input image as normal or cancerous, with an accuracy of 91.33 percent. In the future, the method is tested on a vast amount of data, and new features such as nodule size, texture, and position can be added. Convolutional Neural networks can be used to increase efficiency. Different Stages of lung cancer can be detected along with age and gender implementation.
Acknowledgments

I would like to express my respect and gratitude towards our beloved and respected guide and mentor for his continuous motivation, valuable suggestions guidance, support and technical advice to conduct this research work.

References


