Proc. of Int. Conf. on Intelligence Computing & Information Technology , ICIT

Computer Aided Detection of Lung Cancer using Image Processing Techniques

Brunda S[1] and Rajaram M Gowda[2]
[1]Dept of ISE, Ramaiah Institute of Technology, Bangalore, India
brundas6@gmail.com
[2]Associate Professor, Dept of ISE, Ramaiah Institute of Technology, Bangalore, India
raj.gowda@gmail.com

***Abstract*—In recent years, the image processing techniques are commonly used in medical areas that improves early detection and treatment stages. Time is an important factor in the discovery of any disease in a patient, especially cancer tumors like lung cancer, brain cancer and so on. Lung cancer being the most common cause of death amongst people throughout the world, in this study lung cancer detection system is developed using image processing techniques. Early detection of cancer using image processing technique can enhance patients chance for survival. This paper presents a computer aided classification method, using Computed Tomography (CT) images of lung, based on two classifiers, viz., KNN K-Nearest Neighbor & SVM- Support Vector machine. In this study, preprocessing and segmentation is performed using morphological operations like dilate and erosion. The GLCM extracted features are used for classification process.**

***Index Terms*— Gray Level Co-Occurrence Matrix, Support Vector Machine, K-Nearest Neighbor, Feature Extraction, Segmentation, Classification Technique.**

I. INTRODUCTION

Cancer has become the most serious health issue in the world. The mortality rate of lung cancer is highest among other types of cancers. Survival from lung cancer is directly related to its detection time and growth [5]. Higher chances of successful treatment are possible when the detection is made at the earliest. Cancer is one of the most common cause of death in the US that accounts for nearly 1 of every 4 deaths. Lung cancer can be diagnosed from the CT image of lung in DICOM format. Commonly, the doctor analyses the CT image of lung to detect the occurrence of cancer. The chance of false detection is more likely in manual diagnosis. False detection would be due to the presence of ribs, blood vessels, air in bronchi and so on. So, there is a need to develop a reliable computerized cancer detection system. The best tool for developing such a computerized method is image processing techniques. CT image of lung in DICOM format is processed to diagnose cancerous and non-cancerous tumors.

II. LITERATURE REVIEW

Lung cancer remains the leading cause of cancer-related deaths across the world. Early diagnosis improves the effectiveness of treatment and increase the chance of survival for the patient [7]. Healthy lung tissues

Grenze ID: 02.ICIT.2018.2.503

form darker regions in CT images in contrast to other parts of the chest. Thus an optimum threshold that separates the lungs from all other tissues is to be found. Hu et al. computed iteratively such a threshold to get an initial lung region which is further refined by opening and closing morphological operations. Yim et al. has extracted the lung fields by the region growing method followed by connected-component analysis. Armato et al. used gray-level thresholding to segment the lungs from the thorax. A rolling ball filter was further applied to the segmented lung borders to avoid the loss of juxtapleural nodules. The identified lung fields were used to limit the search space for their lung nodule detection framework. Pu et al. set a threshold to initially segment the lung region which is further refined by segmentation and include juxtapleural nodules. A border marching algorithm was used to march along the lung borders with an adaptive marching step in order to refine convex tracks [11]. Farzad et al. used ensemble of three classifiers including MLP, KNN, SVM using CT images of lung and it showed good results in diagnosing pulmonary nodules. Ziqi et al. applied weighted KNN approach by using weights from feature extraction and optimization.

III. Proposed Methodology

In summary, the system takes lung CT images as an input and applies two major techniques on them comprising image processing techniques and classification technique. Pre-processing of images, lung masking, lung segmentation, background elimination, and feature extraction are done in the first module [4]. In the second module, SVM (Support Vector Machine) and KNN (K-Nearest Neighbor) classifiers are used to classify the lung cancer nodules by using the features, so the lung tumors could be identified.
There are four methods, namely:

- Preprocessing
- Lungs Segmentation using Morphological Operations
- GLCM Feature Extraction
- Classification using KNN and SVM

This section provides the complete description of all the modules of this system. Each module is discussed in detail by mentioning corresponding algorithms.

A. Preprocessing and Lung Segmentation using Morphological Operations

This module provides the complete description of how the input CT scan image is preprocessed to remove unwanted portions and segment the lung regions using morphological operations, to screen for lung cancer.

Algorithm:

Step 1: Read low-dose CT scan image
Step 2: Resize the image to 256x256 to fit the limitation of the resources
Step 3: Convert to gray scale image
Step 4: Convert to binary image
Step 5: Perform Morphological operations
Step 6: Generate the Chest mask
Step 7: Apply the mask on the gray scale image
Step 8: Display Segmented chest region on gray scale image
Step 9: exit

B. GLCM Feature Extraction

This module provides the description of how to calculate the Gray Level Co-occurrence Matrix (GLCM) and the texture features such as contrast, homogeneity, energy and correlation using GLCM.
The GLCM $P_{(d,\theta)}(l_1, l_2)$ represents the probability of occurrence of the pair of gray level (l_1,l_2) separated by a given distance d at angle θ [6].
The GLCM texture measures are defined as follows.

- Energy, the squared elements are summed in the GLCM.

$$F_1 = \sum_{l1=0}^{L-1} \sum_{l2=0}^{L-1} p^2(l_1, l_2) \qquad (1)$$

- Contrast, the area of image is separated in to darkest and brightest area.

$$F_2 = \sum_{k=0}^{L-1} k^2 \sum_{l_1=0}^{L-1} \sum_{l_2=0}^{L-1} p(l_1, l_2) \qquad (2)$$

- Homogeneity, standard or condition to be homogeneous

$$F_5 = \sum_{l_1=0}^{L-1} \sum_{l_2=0}^{L-1} \frac{1}{1 + (l_1 - l_2)^2} p(l_1, l_2) \qquad (3)$$

- Correlation, the interaction with the delayed copy of signal itself as delay of function.

$$F_3 = \frac{1}{\sigma_x \sigma_y} \left[\sum_{l_1=0}^{L-1} \sum_{l_2=0}^{L-1} l_1 l_2 p(l_1, l_2) - \mu_x \mu_y \right] \qquad (4)$$

where μ_x and μ_y are means and σ_x and σ_y are standard deviations of p_x andp_y.

C. Classification using KNN and SVM

This module provides how the GLCM features are used to train the SVM and KNN classifiers to classify the segmented lung regions into cancerous or non-cancerous. The gray level co-occurrence matrix (GLCM) is calculated for the segmented lung regions and features are measured for cancerous images and non-cancerous train images to build the feature vector which is used to train both the classifiers SVM and KNN. Then the GLCM features from the input test image are given to the trained SVM and KNN classifiers and the output is labelled into cancerous or non-cancerous based on the classification result.

We have used different variants of SVM and KNN classifier like Linear Support Vector Machine, Quadratic SVM, Fine Gaussian SVM, Cosine KNN, Medium KNN, Weighted KNN and Fine KNN for providing better performance. Support Vector Machine (SVM) is a supervised machine learning algorithm that is used for classification. It is the coordinates of individual observation; it performs at its best to segregate the two classes using hyperplane. KNN classifier is used for classification and regression. It is easy to interpret the output, calculate time, and predict the power with the help of KNN. K-nearest neighbors is an algorithm which can store the available cases and classifies new cases based on the measure of similarities [11].

Classification includes training and testing stages. During training, the CT scan image of lung is provided as input to the classifier. The feature vector along with the labels of train images are fed to SVM and KNN classifiers. The classifiers are trained to identify cancerous and non-cancerous lung images. In testing stage, unlabeled CT scan images are fed to SVM and KNN classifiers that labels the images as cancerous and non-cancerous accurately using the result.

IV. System Design

Design is an important phase of software development. The design can be known as a creative process where a system organization can be established which satisfies the functional and non-functional system requirements. Large Systems can be decomposed into sub-systems which provides some related set of services. The description of the Software architecture is the output of a design process. In this paper, the system is broken into different modules, with a certain amount of dependency among them.

The system has the following modules:

- Reading the input CT scan and Preprocessing
- Segmentation of lungs using Morphological Operations
- GLCM Feature extraction from lungs region
- Classification of lungs region to cancerous or non-cancerous

The system takes CT scans as the input and performs preprocessing and segmentation of lungs using Morphological operations and detects the presence of cancer in the lungs region.

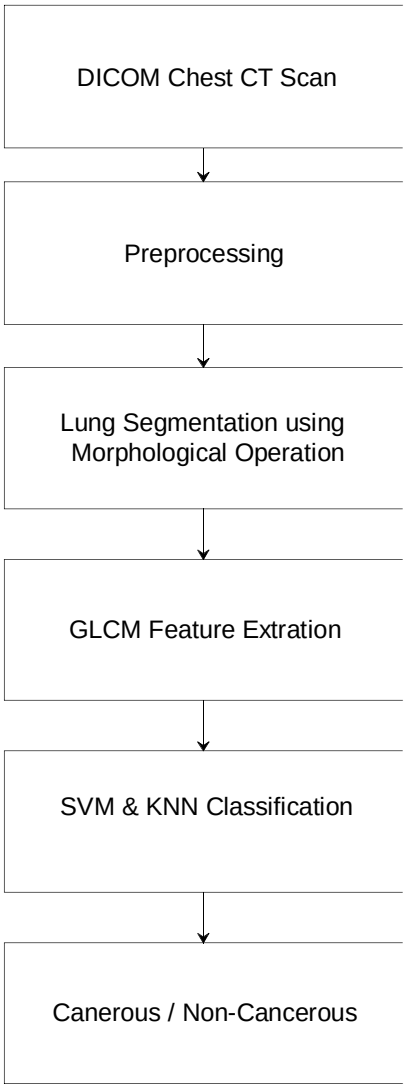

Figure 1. System architecture of lung cancer diagnoses

V. IMPLEMENTATION

A. Datasets

Dataset is obtained from Kaggle website with ground truth information. There are presently 500 cases out of which 289 are cancerous and 211 non-cancerous cases. The dataset contains CT scan images obtained from CT devices which are DICOM series images [1]. Each image has dimension of approximately 525* 525 pixels. Machine learning algorithms are setup and manual dataset for training is provided.

B. Training and Testing

In this study, SVM and KNN is implemented using MATLAB. Manually annotated DICOM images are stored in file system. SVM model and KNN model with pre-defined layer configuration is trained with labelled images. SVM and KNN trained models are saved in .MAT format [1]. About seventy percent dataset is trained and thirty percent of dataset is used for testing. Similar method is followed to diagnose Lung cancer.

VI. Results

The Measures used evaluate performance of algorithms on selected dataset using equations (5), (6), (7).

$$\text{Accuracy} = \frac{\text{Number of correctly classified}}{\text{Total Number of Test Sample}} * 100 \quad (5)$$

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives+False Negative}} * 100 \quad (6)$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative+False Positive}} * 100 \quad (7)$$

To train the SVM and KNN classifiers we have used 70% of the dataset. We have performed 5 fold-classification. We used Linear Support Vector Machine, Quadratic SVM, Fine Gaussian SVM, Cosine KNN, Medium KNN, Weighted KNN and Fine KNN in order to perform classification. SVM and KNN allows a clear separation and classification of the cells into cancerous and non-cancerous with good performance. Thus, it is suitable and reliable for the proposed system [3]. Table 1 summarizes the performance of SVM and KNN classifiers. Fine KNN obtained a high number of TP and TN and reduced the number of FP and FN which lead to successful classification.

TABLE I: PERFORMANCE OF THE SVM CLASSIFIER

System Performance Measurements	Accuracy	Sensitivity	Specificity
Linear SVM	75	76	74
Quadratic SVM	87.5	77	72
Fine Gaussian SVM	79	85	72
Cosine KNN	88	93	82
Medium KNN	87.2	92	82
Weighted KNN	91.9	93	90
Fine KNN	**93.3**	**93**	**93**

VII. Conclusion and Future Work

Experiment is carried on same dataset to extract GLCM features, then train and test on different classifiers like SVM and KNN. The experiment results showed very good accuracy (93.3%), sensitivity (93%), specificity (93%) using Fine KNN among the different variants of KNN. Hence, conclusion of experiment results is -KNN classification performance in lung cancer detection obtains the overall success rate of the system. In this study, the preliminary results of low level classification of SVM and KNN on relatively medium size dataset is obtained. Future scope of this project is to train these algorithms on large dataset and also patient level accuracy on different dataset can be performed. The approach used in project will provide baseline for further researches in lung cancer diagnosis using machine learning algorithm. The proposed method can also be explored in other medical imaging diagnosis such as Breast cancer, Brain tumor Detection and other type of cancers.

References

[1] Akshay M Godkhindi, Rajaram M. Gowda "Automated detection of polyps in CT colonography images using deep learning algorithms in colon cancer diagnosis", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017.

[2] Chaudhary, Anita, and Sonit Sukhraj Singh. "Lung Cancer Detection on CT Images by Using Image Processing", 2012 International Conference on Computing Sciences, 2012.

[3] Fatma Taher, Naoufel Werghi, Hussain Al-Ahmad. "Computer aided diagnosis system for early lung cancer detection", 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), 2015.
[4] Farahani, Farzad Vasheghani, Abbas. Ahmadi, and M.H. Fazel Zarandi. "Lung nodule diagnosis from CT images based on ensemble learning", 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015.
[5] S. Manju Priyanka, R.I. Minu. "Improving the conspicuity of lung nodules by use of " Virtual Dual-Energy" radiography", International Conference on Information Communication and Embedded Systems (ICICES2014), 2014.
[6] Li, Junhua, and Shusen Wang. "An automatic method for mapping inland surface waterbodies with Radarsat-2 imagery", International Journal of Remote Sensing, 2015.
[7] El-Baz, Ayman, Garth M. Beache, Georgy Gimel'farb, Kenji Suzuki, Kazunori Okada, Ahmed Elnakib, Ahmed Soliman, and Behnoush Abdollahi. "Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies", International Journal of Biomedical Imaging, 2013.
[8] K. Manikandan. "Blob based segmentation for lung CT image to improving CAD performance", 2014 International Conference on Recent Trends in Information Technology, 2014.
[9] “Information and Software Technologies", Springer Nature, 2017.
[10] Mahersia, H., M. Zaroug, and L. Gabralla. "Lung Cancer Detection on CT Scan Images: A Review on the Analysis Techniques", International Journal of Advanced Research in Artificial Intelligence, 2015.
[11] https://www.analyticsvidya.com>blog.