IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 04, November 2019)

Machine Learning-based Lung Cancer Image Classification using GLCM and LBP Features

Jaydeep Trivedi
M. Tech. Scholar
Computer Science and Engineering Department
Pacific Institute of Technology
Udaipur, Rajasthan (India)

Ankita Bhargava
Assistant Professor
Computer Science and Engineering Department
Pacific Institute of Technology
Udaipur, Rajasthan (India)

Dr. Prashant Sharma
Associate Professor
Computer Science and Engineering Department
Pacific Institute of Technology
Udaipur, Rajasthan (India)

Abstract – **Most of the models for lung cancer classification based on lung cancer image are various types of the classification model with binarization image pre-processing. This research work proposes a method based on Random forest classifier for lung cancer image classification from the given database images. Feature extraction of the image is accomplished using LBP (Local Binary Pattern) and GLCM (Grey Level Co-occurrence Matrix). Then the extracted features are classified by the Random forest classifier. This work provide the confusion matrix with sensitivity, specificity and accuracy for LBP, GLCM and Hybrid (LBP+GLCM) based approaches.**

Keywords – **GLCM, LBP, Lung Cancer, SCLC, NSCLC.**

I. INTRODUCTION

Lung cancer is the leading cause of cancer death in men and the second in women (after breast cancer) in the countries of the European Union. 85% of patients diagnosed with lung cancer die from the tumor. The incidence of lung cancer increases with age, so as the population ages, it can be anticipated that the number of patients with this tumor will continue to increase.

Lung cancer according to the histopathological classification is divided into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) [1].

The NSCLC accounts for approximately 85% of this type of tumor and the SCLC is 15%. The SCLC grows faster than the non-small cell [2].

Although this characteristic makes the former more susceptible to cytotoxic drugs, it also leads to an earlier development of metastasis. Basic staging systems are different for both large cancer groups and reflect the importance of defining the stage to determine prognosis and treatment [3].

Patients with SCLC and limited disease who do not receive treatment live approximately 3 months from the time of diagnosis. In the case of patients with SCLC and extended disease, it is 2 months [4].

However, in patients with treatment the median survival in patients with limited disease amounts to 16-20 months with a percentage of patients alive at 5 years of 20-25%, while in patients with extended disease is 7-9 months, with less than 5% of patients alive at 5 years. Between 60% and 70% of patients with SCLC have widespread disease at the time of diagnosis.

The short survival times mentioned, reflect the rapid growth and metastasis associated with this histology.

Patients with NSCLC after surgery have a 5-year survival for stage I of 55-65%, for stage II of 40-50% and for stage IIIA of 20-25%. Overall survival at 5 years by adding chemotherapy can increase up to 69% [5].

At the time of diagnosis, only 20-30% of patients with NSCLC have a localized stage. Approximately 50% of patients present advanced stage at the time of diagnosis [6].

The main objective of this paper is to implement a Lung Cancer Image Classification system using Random Forest Classifier. Feature extraction for the database image is done using LBP (Local Binary Pattern) and GLCM (Grey Level Co-occurrence Matrix) approaches. Performance evaluation is done using confusion matrix plot with sensitivity, specificity and accuracy.

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 04, November 2019)

II. PROPOSED METHODOLOGY

A. System Model

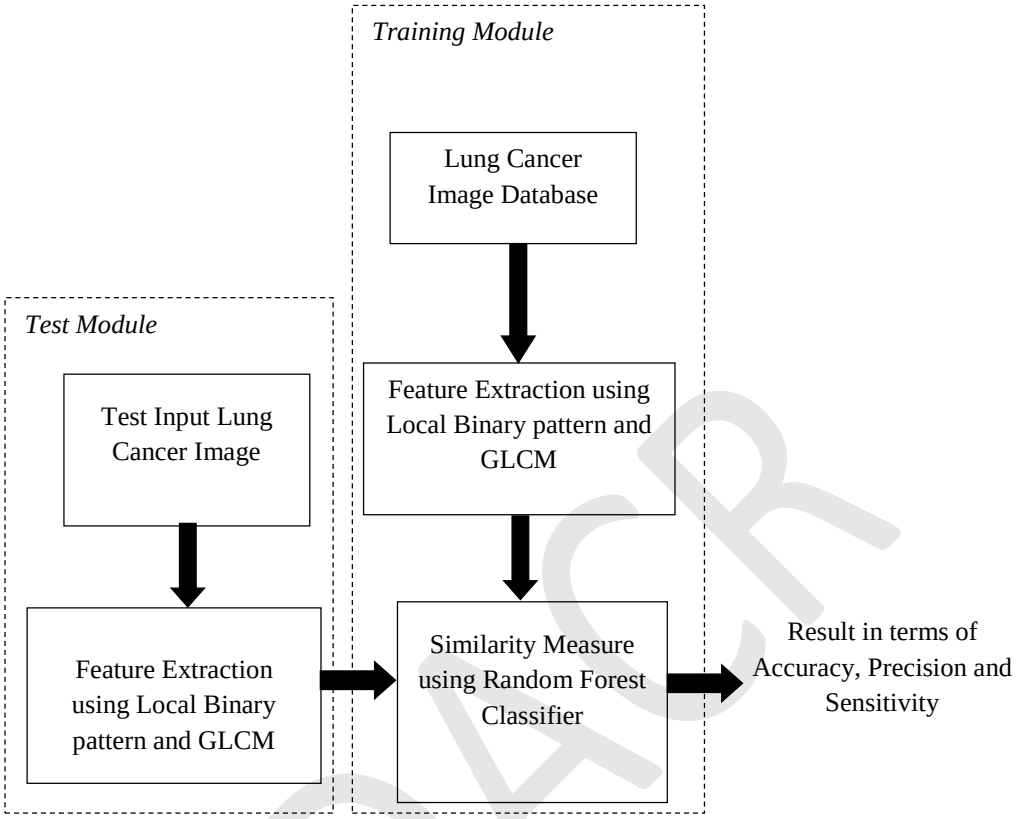

Figure 1: Flow diagram of proposed work

Figure 1 shows the basic block diagram for proposed Lung Cancer Image Classification system. It consists of two modules; training and test. Rest of the methodology is explained as follows:

B. Image Acquisition

There are lung images from Japanese Society Radiology and Technology [7] used in this research work (93 normal lung images and 154 malignant lung images). It is divided into training data and testing data. The input image is of the size of 2048×2048.

C. Feature Extraction

There are two features have been considered for proposed lung cancer classification.

1) Local Binary Pattern

Lung image is separated into small region for computation of LBP for every region image pixel, further histogram of the LBP, is considered as feature vector of lung image. Let us N to form a large histogram representing the image of lung features. Efficacy of the LBP code as a lung index is explained by the fact that the LBP allows to characterize the details of a lung. All non-uniform LBPs are labelled with just a single label when only uniform LBPs are used, while each uniform codes is grouped in a single histogram. For example, when P=8, it has 58 uniform codes but the histogram is of dimension 59. Similarly P=6 produces a histogram of dimension 33.

Given two histograms of $LBP\ H^1, H^2$ of two lungs, the subsequent phase is to use a metric to compute the similarity between these two histograms. In testing the three metrics χ^2, Histogram intersection and Log-likelihood statistic:

$$\chi^2(H^1, H^2) = \sum_i \frac{(H_i^1 - H_i^2)^2}{H_i^1 + H_i^2} \quad (1)$$

2) GLCM (Grey Level Co-occurrence Matrix)

The most commonly used method for mathematically measuring texture is the grey level co-occurrence matrix, or GLCM (Grey Level Co-occurrence Matrix), based on 2nd order statistics. It is a histogram of the grey levels of two dimensions for a pair of pixels (reference pixel and neighbour).

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 04, November 2019)

This matrix approximates the probability of joint distribution of a pair of pixels.
Second Order: are the measures that consider the co-occurrence relation between groups of two pixels of the original image and at a given distance.

Texture Measures
Up to this point we have detailed how a normalized matrix, expressed as probability, is created for a given spatial relationship between two neighbouring pixels. Once constructed, different measurements can be derived from this matrix, in this section some of them are defined, and the measurements whose calculations can be performed manually due to their simplicity are developed in greater depth.
The following is a brief explanation of some textural measures:

Homogeneity:
It is calculated by equation (2).

$$\sum_{i,j=0}^{N-1} P_{i,j} / 1 + (i-j)^2 \qquad (2)$$

Where $P_{i,j}$ the probability of co-occurrence of gray is values i *and* j , for a given distance.
The difference between this GLCM averages the arithmetic mean of the grey values of the window pixels is noted. The mean in the co-occurrence matrix is not simply the average of the original values of the grey levels in the window. The value of the pixel is not weighted by its frequency per se, but by the frequency of its co-occurrence in combination of a certain value of the neighbouring pixel.

Contrast:
It is the opposite of homogeneity, that is, it is a measure of local variation in an image. It has a high value when the region within the scale of the window has a high contrast.

$$\sum_{i,j=0}^{N-1} P_{i,j} (i-j)^2 \qquad (3)$$

Where $P_{i,j}$ the probability of co-occurrence of gray is values i *and* j , for a given distance.
Correlation:

$$\sum_{i,j=0}^{N-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_i^2)}} \right] \qquad (4)$$

The result is between -1 and 1.
As it arises from the equation, this measure is calculated differently from the previous measures, so the information it provides is essentially different, it is independent of the other measures. Therefore it is expected that it can be used in combination with another textural measure.
Some properties of the Correlation are:

- An object has a higher correlation within it than between adjacent objects.
- Nearby pixels are more correlated with each other than more distant pixels.

D. Classification by Random Forest Classifier
The random forest technique modifies the Bagging method applied here to trees by adding a de-correlation criterion between these trees. The idea behind this method is to reduce the correlation without increasing the variance too much. The principle is to randomly choose a subset of variables that will be considered at each level of choice of the best node of the tree.
Consider a training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, a has the number of attributes of the examples of X. Also consider S_t a bootstrap containing m instances obtained by resampling with replacement of S. Let $\{h_1, \ldots h_t\}$ be set of T decision trees. Each tree h_t is built from S_t. For each node of the tree, the partitioning attribute is chosen by considering a number $f(f < a)$ of randomly selected attributes (among the attributes a). To classify a new instance x, the random forest classifier performs a uniformly weighted majority vote of classifiers in that set for instance x. The algorithm illustrates this principle [8].

Algorithm:
Input: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, *the training set.*
Input: T, *the number of decision trees in the random forest.*
For $t = 1, \ldots, T$ *do*
1. *Generate a Bootstrap sample* S_t *of size* m *from* S
2. *Create a decision tree* h_t *from* S_t *by recursively repeating for each node of the tree the following steps:*
   a. *Randomly select* f *attributes among* a *attributes.*
   b. *Choose the partitioning attribute among* f
   c. *Partition the node into two child nodes*

End for
Output: H, *the random forest classifier*

III. Simulation And Results

The performance of proposed algorithms has been studied by means of MATLAB simulation.

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 04, November 2019)

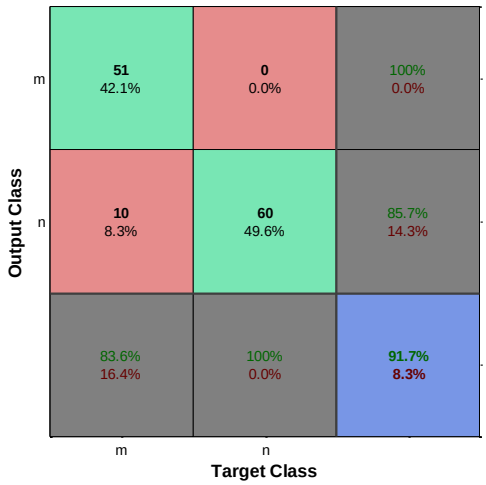

Figure 2: Confusion matrix plot for LBP based approach

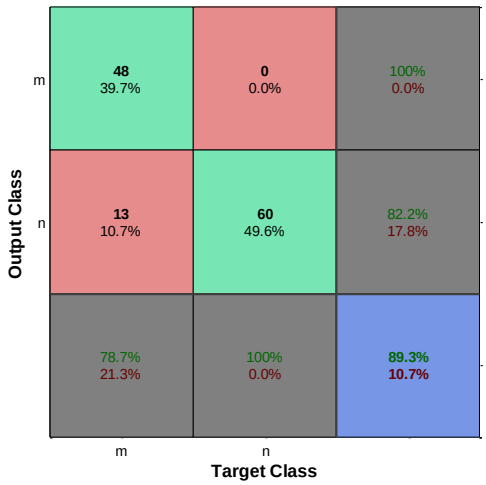

Figure 3: Confusion matrix plot for GLCM based approach

Figure 4: Confusion matrix plot for GLCM-LBP based Hybrid Approach

IV. Conclusion

Lung cancer is one kind of dangerous diseases, so it is necessary to detect early stages. But the detection of lung cancer is most difficult task. From the literature review many techniques are used for the detection of lung cancer but they have some limitations. In the proposed method in which first step is image acquisition, and then feature extraction, and then these features are classified by the random forest classifier. The proposed system successfully detects the lung cancer from CT scan images. It can be said that the system achieve its desired expectation. The proposed system test 121 types of lung CT images and obtains the result where overall success rate of the system is 95.9% which meet the expectation of system.

Reference

[1] Molina, Julian R., Ping Yang, Stephen D. Cassivi, Steven E. Schild, and Alex A. Adjei. "Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship." In Mayo Clinic Proceedings, vol. 83, no. 5, pp. 584-594. Elsevier, 2008.

[2] Cerfolio, Robert James, Ayesha S. Bryant, Buddhiwardhan Ojha, and Mohammad Eloubeidi. "Improving the inaccuracies of clinical staging of patients with NSCLC: a prospective trial." The Annals of thoracic surgery 80, no. 4 (2005): 1207-1214.

[3] Shi, Yuankai, Joseph Siu-Kie Au, Sumitra Thongprasert, Sankar Srinivasan, Chun-Ming Tsai, Mai Trong Khoa, Karin Heeroma, Yohji Itoh, Gerardo Cornelio, and Pan-Chyr Yang. "A prospective, molecular epidemiology study of EGFR mutations in Asian patients with advanced non–small-cell lung cancer of adenocarcinoma histology (PIONEER)." Journal of thoracic oncology 9, no. 2 (2014): 154-162.

[4] Miah, Md Badrul Alam, and Mohammad Abu Yousuf. "Detection of lung cancer from CT image using image processing and neural network." In 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), pp. 1-6. 2015.

[5] Tiwari, Arvind Kumar. "Prediction of lung cancer using image processing techniques: a review." Advanced Computational Intelligence: An International Journal (ACII) 3, no. 1 (2016).

[6] Nurtiyasari, Devi, and Dedi Rosadi. "The application of Wavelet Recurrent Neural Network for lung cancer classification." In Science and Technology-Computer (ICST), 2017 3rd International Conference on, pp. 127-130. IEEE, 2017.

[7] Shiraishi, Junji, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules." American Journal of Roentgenology 174, no. 1 (2000): 71-74.

[8] Breiman, L. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA, 2002.