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 02, September 2019)

Lung Cancer Image Classification system using Random Forest Classifier

Sudeep Gujar
M. Tech. Scholar
Department of Computer Science
Shiv Kumar Singh Institute of Science and Technology, Indore, M. P. (India)

Laal Singh Chauhan
HOD, Department of Computer Science
Shiv Kumar Singh Institute of Science and Technology, Indore, M. P. (India)

Nisha Kumawat
Assistant Professor
Department of Computer Science
Shiv Kumar Singh Institute of Science and Technology, Indore, M. P. (India)

Abstract – **Lung cancer is a real public health problem. Indeed, it is the leading cause of cancer mortality in the world, survival at 5 years is only 15% and this is largely due to late diagnosis and a high metastatic power. Improving the management of this type of cancer therefore implies a better knowledge of the processes of oncogenesis and tumor invasion. 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 paper 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 Gabor Wavelet and GLCM (Grey Level Co-occurrence Matrix). Then the extracted features are classified by the Random forest classifier. This paper provides the confusion matrix with sensitivity, specificity and accuracy for Gabor wavelet, GLCM and Hybrid (Gabor + GLCM) based approaches.**

Keywords – **Lung Cancer, Gabor Wavelet, GLCM, Random Forest Classifier.**

I. INTRODUCTION

Cancer as a disease is well known throughout the history of mankind. This entity, through a process of carcinogenesis involves different genetic mutations and epigenetic changes in protooncogenes, tumor suppressor genes, cell repair genes and microRNAs, in order to confer a malignant phenotype to a cell clone; that is, it acquires the ability to be self-dependent, invade, evade the immune response and metastasize to other parts of the body. These genetic changes are caused by environmental, physical and biological exposure, which increases the susceptibility to cancer and modifies the epidemiological profile of each country, explaining the great variability in the incidence of morbidity and mortality worldwide due to cancer.

Pulmonary carcinoma was considered until the middle of the last century as a rare disease. Since 1930 its frequency has increased and is currently the most frequent malignant tumor in the world. Several authors denote an increase in the frequency of lung cancer in recent decades. Currently in our country it is among the three leading causes of death from malignant tumors in adults over 35 years of age and is more frequent in men, although a worldwide increase in cases of women has been reported. Its incidence is very high and due to its lethality, the mortality figure is very close to the incidence and the latter is expected to increase in the years [1].

Lung cancer is a real public health problem. Indeed, it is the leading cause of cancer mortality in the world and the 5-year survival is 15%. This is largely due to late diagnosis and high metastatic potential [2]. Indeed, 60% of bronchial cancers are diagnosed at a metastatic stage. In addition, metastases appear very rapidly after the appearance of primary tumors. Improving the management of this type of cancer therefore implies a better knowledge of the processes of oncogenesis and tumor invasion [3].

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 [4].

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 Gabor Wavelet and

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 02, September 2019)

GLCM (Grey Level Co-occurrence Matrix) approaches. Performance evaluation is done using confusion matrix plot with sensitivity, specificity and accuracy.

II. PROPOSED METHOD

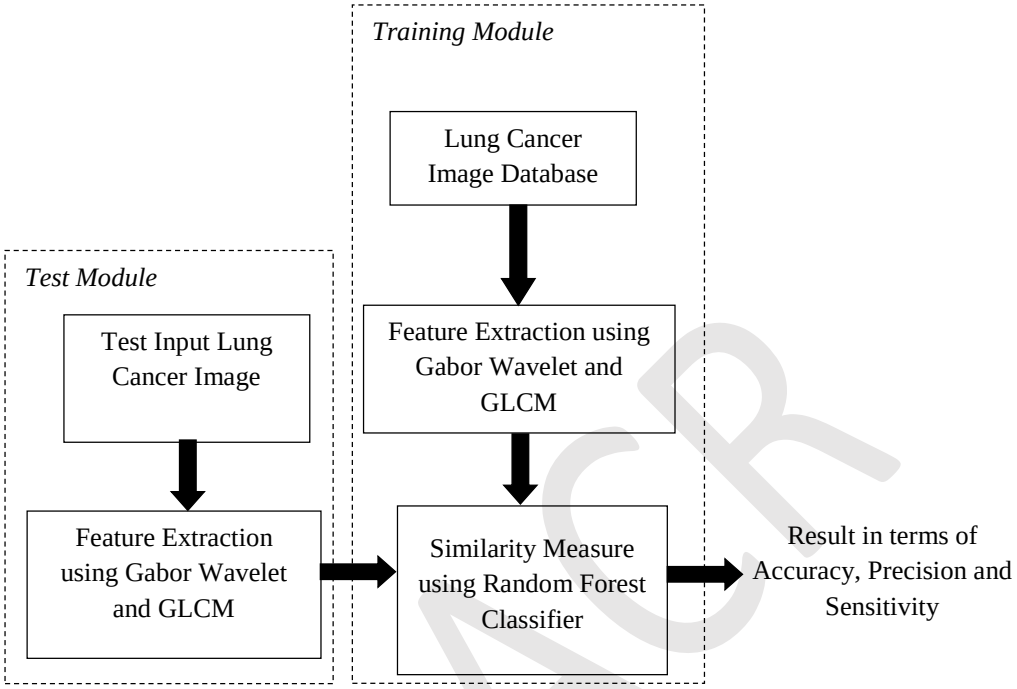

Figure 1: Flow diagram of proposed work

A. *Image Acquisition*

There are lung images from Japanese Society Radiology and Technology [5] used in this paper (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.

B. *Feature Extraction*

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

1) Gabor Wavelet

Gabor's Eye, developed by Dennis Gabor, is extensively used as a treatment of images because the Gabor wavelets salient properties: the localization frequency and selectivity in orientation. Frequency representations and orientation of Gabor are according to biometric recognition system. The Gaussian envelope for lung cancer classification is represented as follows:

$$\psi_{u,v}(z) = \frac{\|K_{u,v}\|^2}{\sigma^2} e^{\frac{\|k_{u,v}\|^2 \|z\|^2}{2\sigma^2}} \left[e^{ik_{u,v}z} - e^{-\frac{\sigma^2}{2}} \right] \quad (1)$$

Where $z = (x; y)$ is the coordinate point $(x; y)$. Where u and v are orientation and frequency respectively for kernels of Gabor. $\|.\|$ is the standard operator and σ is standard deviation of the Gaussian envelope.

The Gabor wavelet is the representation of convolution product of frequency and orientation claimed from equation (1). The convolution of image I and of a kernel of Gabor $\psi_{u,v}(z)$ is defined by:

$$G_{u,v}(z) = I(z) * \psi_{u,v}(z) \quad (2)$$

The interest of using Gabor Wavelet to extract database image features is capturing the information in orientations and resolutions. In addition, they are invariant of illumination, distortions and variations in scale. Therefore, if only the amplitude response is considered, "Jet" and it has been widely used in the oldest systems, such as the DLA and the EGBM. Note that these are methods based on the characteristic points which must be detected very precisely. Several metrics have been tested for characteristics based on Gabor and the one that is most often used is the cosine distance.

2) GLCM (Grey Level Co-occurrence Matrix)

The following is a brief explanation of some textural measures:

Homogeneity: It is calculated by equation (3).

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 02, September 2019)

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

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.

$$\sum_{i,j=0}^{N-1} P_{i,j} (i-j)^2 \quad (4)$$

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

Correlation:

$$\sum_{i,j=0}^{N-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_i^2)}} \right] \quad (5)$$

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.

C. Classification by Random Forest Classifier

The above extracted feature values can be combined to get optimal dataset for training of the classifier.

A random forest is a classifier containing of one group of structured tree predictors $[T(x,\ominus_k), k = 1, \ldots.]$ where the $[\ominus_k]$ are random vectors of identical distributions and where every tree provides a unit poll for the furthermost common class of each entry x.

The main advantage of this structure is that it avoids the danger of over-learning for any method of prediction based on induction. BREIMAN [6] shows that when the number of trees involved in the prediction forest increases, the generalization error rate converges to a limit value, of which an upper bound can be estimated on the basis of the characteristics intrinsic features of the forest.

The classification trees in RF is built by selecting features from random samples to obtain a class label.

The organizational system of the RF classifier is depicted in Figure 2. The RF classifier is formed by a number of base learners and each base learner acts as an independent binary tree which adapts recursive partitioning.

The best feature is selected by Gini index, which is used to build the binary tree. It has the following advantages:

- RF is one of the most accurate classifier in present scenario.
- Overfitting is reduced due overgrowing the trees hence it is ease to handle.
- It accepts a large number of input variables without any deletion of the variables.
- The number of base learners is the only setting parameter to give the highest accuracy.

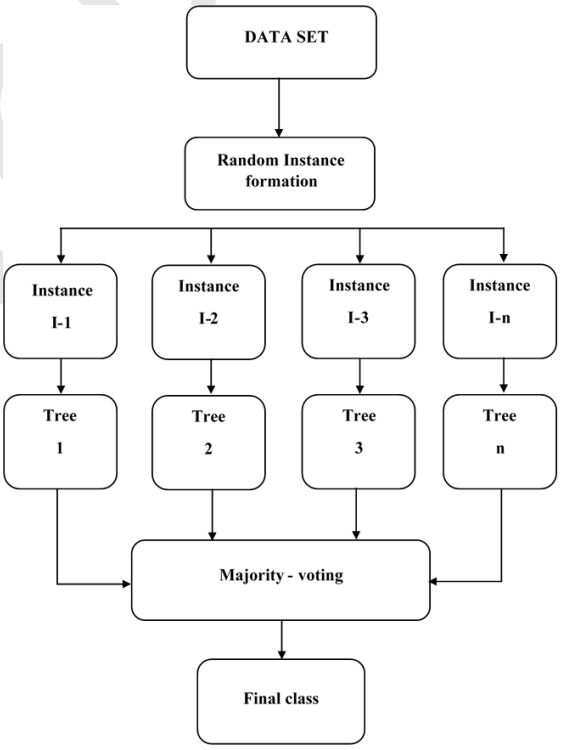

Figure 2: Random forest classifier structural network [6]

If the marginal function of a random forest $T(X,\ominus)$

$$mr(X,Y) = P_{\ominus}(T(X,\ominus) = Y) - \max_{j \neq T} P_{\ominus}(T(X,\ominus) = j) \quad (6)$$

Which represents the confidence level of the ranking established by the trees of this forest on the population (X, Y), measured by the difference of probability between the prediction of the correct class Y and the best class erroneous j ≠ Y, one can

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 02, September 2019)

define the prediction value of a game of trees {T (x, Θ)} by the mathematical expectation of this function

$$s = E_{x,y}[mr(X,Y)] \quad (7)$$

The dependency between trees in a forest ρ (⊖,⊖ ') is measured by the correlation between their gross marginal functions, and it is evaluated for fixed and distinct parameter values ⊖,⊖. By means of these definitions, an upper limit to the error in generalization (TEG) of any random forest is given by the relation.

$$TEG \leq \bar{\rho}(1-s^2)/s^2 \quad (8)$$

The tree structured classifiers combine to form a random forest classifier (Figure 3 and Figure 4) [6].

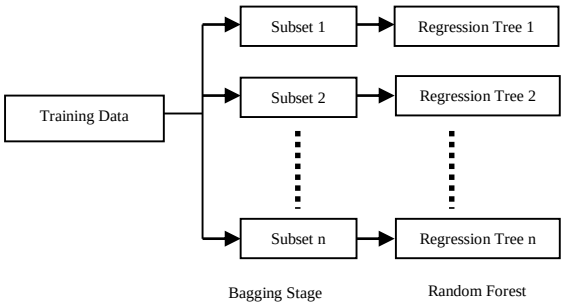

Figure 3: Algorithmic Sequence of Random Forest Classifier Training Phase

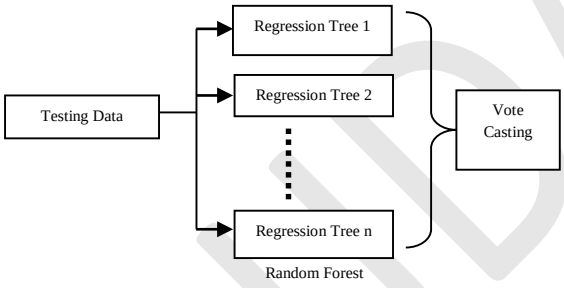

Figure 4: Algorithmic Sequence of Random Forest Classifier Testing Phase

The input objects are sourced to every tree and individual unit vector is voted by every tree. The node from individual classes with maximum number of votes is given as output class by the forest. More than just prediction the random forest can be trained to achieve information on data proximities for feature generation. The variables closer to responsible variables are predicted by their importance while their relation is the function of partial independencies.

In training of N cases, the trees are trained for sampling [6]. This sampling is a random process that replaces the original data with the bootstrap sample. Out of M number of input variables, m variables are selected in random manner ($m \ll M$), and best split on these predictors split the node. During the entire operation of training, the value of m is hold for constant. Generally the value of m is selected M times smaller than M inputs. Without pruning, each tree grows to the maximum possible range. This training generates multiple numbers of trees with the maximum value decided byN_{tree}. The length of tree roots (depth of tree) is estimated via parameter node size (i.e. no. of leaf nodes) that is generally fixed to unity. In testing phase the input is fed to forest that runs to all the trees and classification from each individual is recorded as vote. The instance that gets maximum no. of votes is declared as winner or output.

III. Simulation And Results

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

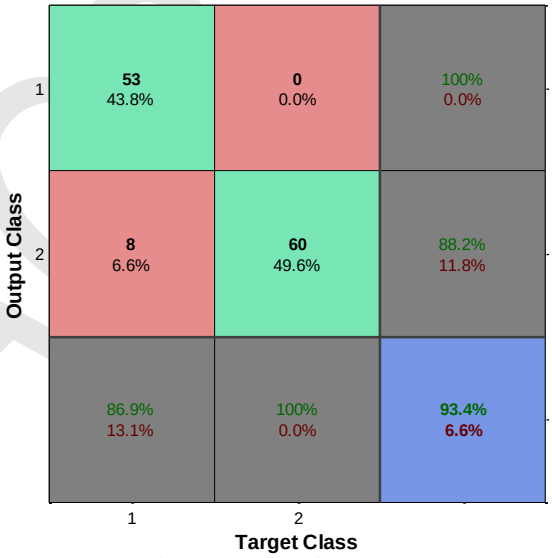

Figure 5: Confusion matrix plot for Gabor wavelet based approach

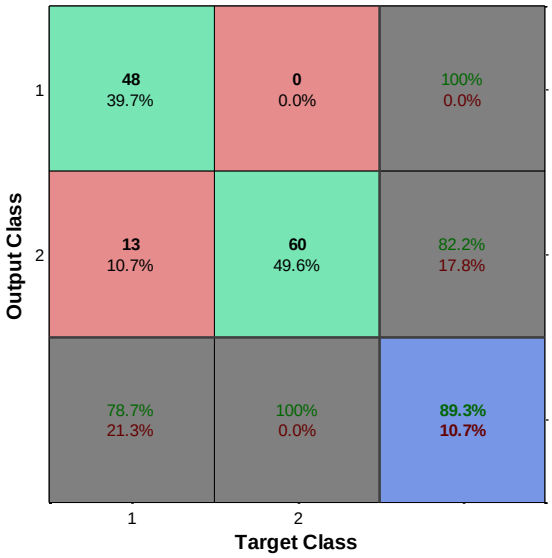

Figure 6: Confusion matrix plot for GLCM based approach

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 02, September 2019)

Figure 7: Confusion matrix plot for GLCM-Gabor wavelet 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% 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] Yu, Tian, John Bachman, and Zhi-Chun Lai. "Mutation analysis of large tumor suppressor genes LATS1 and LATS2 supports a tumor suppressor role in human cancer." *Protein & cell* 6, no. 1 (2015): 6-11.

[3] Goldstraw, Peter, Kari Chansky, John Crowley, Ramon Rami-Porta, Hisao Asamura, Wilfried EE Eberhardt, Andrew G. Nicholson et al. "The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer." *Journal of Thoracic Oncology* 11, no. 1 (2016): 39-51.

[4] Thompson, Elizabeth D., Edward B. Stelow, Stacey E. Mills, William H. Westra, and Justin A. Bishop. "Large cell neuroendocrine carcinoma of the head and neck: a clinicopathologic series of 10 cases with an emphasis on HPV status." *The American journal of surgical pathology* 40, no. 4 (2016): 471.

[5] 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.

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