Eurasian Journal of Analytical Chemistry
ISSN: 1306-3057 OPEN ACCESS 2018 13 (2): 441-446

Automatic Classification of Lung Tumors Using KNN, SVM and CNN

Hema Rajini Narayanan

Received: 01 March 2018 ▪ Revised: 10 April 2018 ▪ Accepted: 24 April 2018

Abstract: A lung tumor classification system has been designed and developed. This work presents a new approach to the automated classification of lung tumors based on texture features, which separate lung tumor images from healthy tissues in computed tomography images. The feature image used for the tumor classification consists of computed tomography images. The application of the proposed method for tracking tumor is demonstrated to help pathologists distinguish its type of lung tumor. A classification with an accuracy of 89%, 97% and 98%, has been obtained by k-nearest neighbor, support vector machine and convolutional neural networks.

Keywords: Support Vector Machine**,** k-nearest Neighbor, Convolutional Neural Networks, Computed Tomography.

INTRODUCTION

Computed Tomography (CT) has outperformed conventional radiography in the screening of lungs because it generates very detailed high-resolution images and can show early-stage lesions that are too small to be detected by conventional X-ray. CT has been widely used to detect numerous lung diseases, including pneumoconiosis, pneumonia, pulmonary edema, and lung cancer. Early detection of diseases is very crucial for treatment planning. However, it is considered one of the most challenging tasks performed by radiologists due to the huge amount of data generated by CT scan. Therefore, computer-aided diagnostic (CAD) systems are needed to assist radiologists in the analysis and evaluation of CT scans.

In last decades it was very difficult to diagnose lung tumor at early stages with the help of image processing and pattern recognition. But with the time new hybrid lung tumor detection method come into existence and accuracy of diagnosis also has got improved. Many methods were proposed that were based on the subtraction between two serial mass chest radiography. This was proposed to detect new lung nodules.

A CAD system analyzes medical images in several steps: first a preprocessing step for noise reduction and enhancing the image quality and then segmentation step is used to differentiate region of interest (ROI). This step is used to differentiate the tumor region from healthy tissues. After segmentation, different features such as geometrical, textural, and statistical features are extracted. Finally, a classification or evaluation step is done to evaluate and diagnose the ROI based on extracted features.

There are many computer-aided classification systems for pulmonary nodules in lung images in the literature, most of them are used to detect and classify abnormalities. In the earlier work the classification of lung tumor includes the work of Patil et al. and Kuruvilla et al., they used artificial neural networks to classify lung cancer images based on the features extracted from lung segmented images [1-2]. Nevertheless, Patil used geometrical features for classification and achieved only 83% accuracy of classification. And Kuruvilla proposed statistical parameters as features for classification and achieved accuracy of 93.3%. Another work by Depeursinge et al., classified different lung tissue patterns using discrete wavelet frames combined with gray-level histogram features [3]. However, the main limitation of this work was the lack of resolution in scales with the decomposition, along with required feature weighting while merging features from different origins.

Aggarwal et al., proposed a model that provides classification between nodules and normal lung anatomy structure [4]. The method extracts geometrical, statistical and gray level characteristics. LDA is used as classifier and optimal thresholding for segmentation. The system achieves 84% accuracy, 97.14%

Hema Rajini Narayanan, Department of Computer Science and Engineering, Alagappa Chettiar Government College of Engineering and Technology, Karaikudi, Tamilnadu, India. E-mail: auhemasmith@yahoo.co.in

sensitivity and 53.33% specificity. Although the system detects the cancer nodule, its accuracy is still unacceptable. Therefore, combination of any of its steps in this model does not provide probability of improvement. Jin et al., used convolution neural network as classifier in his CAD system to detect the lung cancer [5]. The system achieves 84.6% of accuracy, 82.5% of sensitivity and 86.7% of specificity. The advantage of this model is that it uses circular filter in ROI extraction phase which reduces the cost of training and recognition steps. Although, implementation cost is reduced, it has still unsatisfactory accuracy. Sangamithraa et al., uses K-mean unsupervised learning algorithm for segmentation. It groups the pixel dataset according to certain characteristics [6]. For classification this model implements back propagation network. Features like entropy, correlation, homogeneity, PSNR, SSIM are extracted using gray-level co-occurrence matrix (GLCM) method. The system has accuracy of about 90.7%. To perform preprocessing median filter is used for noise removal which can be useful for this new model to remove the noise and improve the accuracy.

Rendon Gonzalez et al., proposed a system that classifies lung cancer as benign or malignant [7]. The system uses the priori information and Housefield Unit to calculate ROI. Shape features like area, eccentricity, circularity, fractal dimension and textural features like mean, variance, energy, entropy, skewness, contrast, and smoothness are extracted to train and classify the support vector machine to identify whether the nodule is benign or malignant. The advantage of this model is that it classifies cancer as benign or malignant, however the limitation of it is that prior information is required about region of interest.

The rest of this paper is organized as follows. Section 1 presents the introduction as well as the studies of several research papers are portrayed. Section 2 presents the proposed technique, utilized in this work for classification of lung images. In this section, pre-processing, feature extraction and classification are presented. Section 3 experimentally demonstrates the performance of the proposed method. Finally, Section 4 describes the conclusion of this paper.

MATERIALS AND METHOD

Materials

For validating the results of the presented model, a benchmark image database is employed which comprises of 50 low-dosage and stored lung CT images. Each image is 1.25mm slice thickness and is obtained in a single breath.

Proposed Method

This proposed research work is used to improve lung tumor classification in lung CT images. The proposed method has three stages, namely pre-processing, feature extraction and classification. The proposed technique for automatic CT lung tumor image classification is illustrated in Figure 1. The proposed system is developed using Matlab (The Math Works, Inc., Natick, MA, USA). In the first stage, noise is suppressed using an image filtering. In the second stage gray level co-occurrence matrix-based features are extracted. Finally, KNN, SVM and CNN classifiers are used to classify the type of lung tumor images.

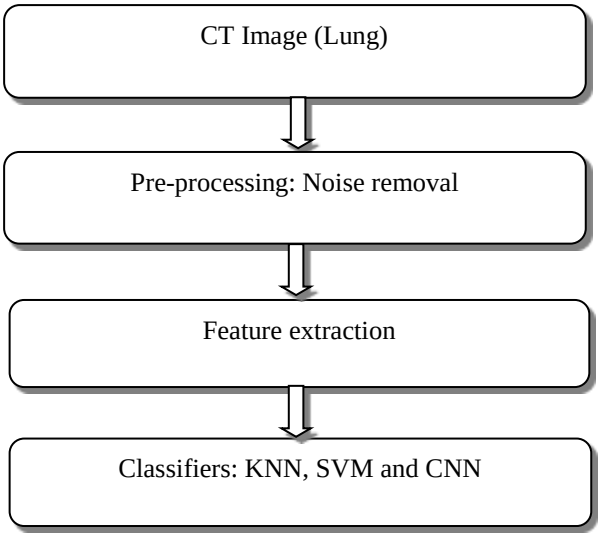

Figure 1: Methodology of the Proposed Technique

a) Pre-processing

To reduce noise, median filtering using a 3-by-3 square kernel is applied [8]. Median filter is chosen because it is less sensitive to extreme values and able to remove outliers without reducing sharpness of the image. This produces a more homogeneous background in which abnormalities become more conspicuous.

b) Texture Features from Gray Level Co-occurrence Matrix

Texture is a repeating pattern of local variations in image intensity. The co-occurrence matrix is a statistical method used for texture analysis. As the name suggests, the co-occurrence matrix is constructed from the image by estimating the pair wise statistics of pixel intensity. The use of the co-occurrence matrix is based on the hypotheses that the same gray-level configuration is repeated in a texture. This pattern will vary more by fine textures than by coarse textures. The co-occurrence matrix *P(i, j|d, θ)* counts the co-occurrence of pixels with gray values *i* and *j* at a given distance *d* and in a given direction *θ.* According to the number of intensity points (pixels) in each combination, statistics are classified into first-order, second-order and higher-order statistics. In the first order, texture measures are statistics calculated from an individual pixel and do not consider pixel neighbor relationships. The gray level co-occurrence matrix (GLCM) method is a way of extracting second order statistical texture features [9]. However, the performance of a given GLCM based feature, as well as the ranking of the texture features; depend on the number of gray levels used. The following notations are μ is the mean value of P. μ_x, μ_y, σ_x and σ_y are the means and standard deviations of P_x and P_y. G is the size of the co-occurrence matrix. Here the number of rows and columns of the co-occurrence matrix is equal. The following GLCM based texture features are extracted in this research work: contrast, correlation, energy, homogeneity entropy, angular second moment and inverse difference moment. They are defined in eqs. (1)-(7).

Contrast

$$Contrast = \sum_{n=0}^{G-1} n^2 \left\{ \sum_{i=1}^{G} \sum_{j=1}^{G} P(i,j) \right\}, |i-j| = n \tag{1}$$

Contrast is a measure of the local variations present in an image. This measure of contrast favors contributions from *P (i, j)* away from the diagonal, i.e. *i= j*. If there is a large amount of variations in an image, the *P[i,j]*'s will be concentrated away from the main diagonal and the contrast will have a high value.

Correlation

$$Correlation = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{\{i \times j\} \times P(i,j) - \{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y} \tag{2}$$

Correlation is a measure of gray level linear dependence between the pixels at the specified positions relative to each other. The correlation will be higher if an image contains a considerable amount of linear structure.

Energy

$$Energy = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p(i,j)^2 \tag{3}$$

The energy of a texture describes the uniformity of the texture. Energy is 1 for a constant image.

Homogeneity

$$Homogeneity = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{P(i,j)}{1+|i-j|} \tag{4}$$

Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. Homogeneity is 1 for a diagonal GLCM. A homogeneous image will result in a co-occurrence matrix with a combination of high and low *P [i,j]*'s. A heterogeneous image will result in an even spread of *P [i,j]*'s.

Entropy

$$Entropy = -\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} P(i,j) \times log\big(P(i,j)\big) \quad (5)$$

Entropy statistic measures the disorder or complexity of an image. Complex textures tend to have high entropy. Entropy is strongly, but inversely correlated to energy.

Angular Second Moment (ASM)

$$ASM = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\{P(i,j)\}^2 \quad (6)$$

ASM is a measure of homogeneity of the image. A homogeneous image contains only a few gray levels, GLCM gives only a few but relatively high values of *P(i,j)*. Thus, the sum of squares also will be high.

Inverse Difference Moment (IDM)

$$IDM = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} \frac{1}{1+(i-j)^2} P(i,j) \quad (7)$$

IDM is also influenced by the homogeneity of the image. Because of the weighting factor $(1+(i-j)^2)^{-1}$ IDM will get small contributions from inhomogeneous areas *(i j)*. The result is a low IDM value for inhomogeneous images and higher value for homogeneous images.

c) Classifiers: KNN, SVM and CNN

Classification is the process of classifying the given input by training with a suitable classifier. Deep learning and Support Vector Machine (SVM) classifiers are the best classifiers suggested by many researchers which can be opted for the lung tumor classification of CT images. It is independent of dimensionality and feature space. Convolutional Neural Networks (CNN) is one of the most remarkable approaches of deep learning, in which multiple layers of neurons are formed in a robust manner. In this work simple classifier like KNN is also used.

k-Nearest neighbor classifier: One of the simplest classification techniques is the k-nearest neighbor (k-NN) classifier. Classification of the input feature vector X is done by determining the k closest training vectors according to a suitable distance metric. Vector X is then assigned to that class to which the majority of that k-nearest neighbors belong. The k-NN algorithm is based on a distance function and a voting function in k-nearest neighbors; the metric employed is the Euclidean distance measure [10]. The k-NN classifier is a conventional nonparametric supervised classifier that is said to yield good performance for optimal values of k. Like most learning algorithms, k-NN algorithm consists of a training phase and a testing phase. Data points are given in an n-dimensional space in the training phase. The labels associated with the data points designate their class in the training phase. In the testing phase, unlabeled data are given and the algorithm generates the list of the k-nearest (already classified) data points to the unlabeled point. This classifier returns the class of the majority of that list.

Support vector machine (SVM) is a powerful supervised classifier and accurate learning technique. From the statistical theory it was derived and developed by Vapnick in 1982. It yields successful classification results in various application domains, e.g. medical diagnosis. SVM is based on the structural risk minimization principle from the statistical learning theory [11]. The kernel controls the empirical risk and classification capacity in order to maximize the margin between the classes and minimize the true costs. SVM searches an optimal separating hyper-plane between members and non-members of a given class in a higher dimensional feature space. The inputs to the SVM algorithm are the features extracted using the GLCM method. In our method, two classes are benign or malignant lung tumors.

Convolutional neural networks: CNNs achieve better classification accuracy on large scale datasets due to their capability of joint feature and classifier learning [12]. The convolutional layer plays a vital role in the operation of CNN. The layers parameters focus around the use of learnable kernels. These kernels are usually small in spatial dimensionality, but spreads along the entirety of the depth of the input. When the data hits a convolutional layer, the layer convolves each filter across the spatial dimensionality of the input to produce a 2D activation map.

The fully-connected layer contains neurons of which are directly connected to the neurons in the two adjacent layers, without being connected to any layers within them. This is analogous to way that neurons

are arranged in traditional forms of ANN. Pooling layers reduce the dimensionality of the representation, and thus further reduce the number of parameters and the computational complexity of the model.

RESULTS AND DISCUSSION

To validating the results of the proposed method, a benchmark image database is employed which comprises of 50 low-dosage and stored lung CT images. Each image is 1.25mm slice thickness and is obtained in a single breath. This section portrays some experimental results on lung CT images. It has tested on dataset of lung images consisting of benign and malignant lung tumor images. The abnormal lung image set consists of images of lung affected by a lung lesion.

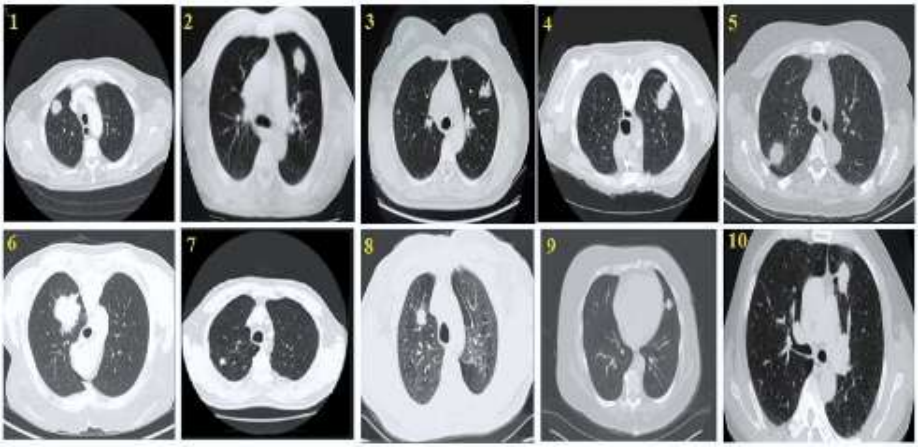

Figure 2: Input CT lung images

In the first stage, noise is suppressed using an image filtering. In the second stage, seven texture features are extracted using gray level co-occurrence matrix. In our research work, seven features are extracted. They are contrast, correlation, energy, homogeneity, entropy, angular second moment and inverse difference moment. Finally, KNN, SVM and CNN classifiers are used to classify the type of tumor images. Figure 2 shows the input lung tumor images.

To evaluate the performance of the classifiers in terms of sensitivity (also called recall in some fields), specificity and accuracy. The formulae for these are given in eqs. (8)-(10). The three terms are defined as follows: Sensitivity (true positive fraction) is the probability that a diagnostic test is positive and it states that the person has the tumor disease, Specificity (true negative fraction) is the probability that a diagnostic test is negative and that the person does not have the disease.

$$Recall\ or\ Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

Accuracy is the probability that a diagnostic test is correctly performed.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

A classification with an accuracy of *89%, 97% and 98%* has been obtained by, k-nearest neighbor, support vector machine and convolutional neural networks. The classification performance of the proposed method is tested at slice level. The performance of our algorithm is excellent. The application of the proposed method for tracking tumor is demonstrated to help pathologists distinguish its type of tumor. Table 1 presents the performance of classifiers at the slice level.

Table 1: Classification Accuracy for the Used Classifiers

Classifier	KNN	SVM	CNN
Accuracy	89%	97%	98%

CONCLUSIONS

Thus, an automated method for classification of two types of tumors in lung CT images based on gray level co-occurrence matrix is developed. This system has been successfully tested on large lung images causing lung tumor. The proposed system helps the physicians to know about the type of lung tumors, for further treatment. A classification with an accuracy of *89%, 97% and 98%* has been obtained by, k-nearest neighbor, support vector machine and convolutional neural networks. The system can be designed to classify other types of cancers as well with few modifications.

REFERENCES

[1] Patil, S.A., & Kuchanur, M.B. (2012). Lung cancer classification using image processing. *International Journal of Engineering and Innovative Technology*, *2*(3), 37-42.

[2] Kuruvilla, J., & Gunavathi, K. (2014). Lung cancer classification using neural networks for CT images. *Computer methods and programs in biomedicine*, *113*(1), 202-209.

[3] Depeursinge, A., Sage, D., Hidki, A., Platon, A., Poletti, P. A., Unser, M., & Muller, H. (2007, August). Lung tissue classification using wavelet frames. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 6259-6262.

[4] Aggarwal, T., Furqan, A., & Kalra, K. (2015). Feature extraction and LDA based classification of lung nodules in chest CT scan images. *In International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1189-1193.

[5] Jin, X. Y., Zhang, Y. C., & Jin, Q. L. (2016). Pulmonary nodule detection based on CT images using convolution neural network. In *9th International symposium on computational intelligence and design (ISCID)*, *1*, 202-204.

[6] Sangamithraa, P. B., & Govindaraju, S. (2016). Lung tumour detection and classification using EK-Mean clustering. In *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2201-2206.

[7] Rendon-Gonzalez, E., & Ponomaryov, V. (2016). Automatic Lung nodule segmentation and classification in CT images based on SVM. In *2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW)*, 1-4.

[8] Zarandi, M. F., Zarinbal, M., & Izadi, M. (2011). Systematic image processing for diagnosing brain tumors: A Type-II fuzzy expert system approach. *Applied soft computing*, *11*(1), 285-294.

[9] Haralick, R.M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics, SMC, 3*(6), 610-621.

[10] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

[11] Cristianini, N., & Shawe Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. *First Edition, Cambridge University Press, England.*

[12] Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015). Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, 588-599.