EARLIER DETECTION OF LUNG CANCER AND ITS RECURRENCE RATE BY GENOMIC BIOMARKERS

SUMAIYA PARVEEN.K, II- M.E. (EMBEDDED SYSTEM TECHNOLOGIES), DEPT OF ECE, VIVEKANANDHA INSTITUTE OF ENGINEERING AND TECHNOLOGY FOR WOMEN, TIRUCHENGODE, NAMAKKAL TAMILNADU, INDIA
Prof. M.NITHYA M.E.,ASSISTANT PROFESSOR (ECE), VIVEKANANDHA INSTITUTE OF ENGINEERING AND TECHNOLOGY FOR WOMEN, TIRUCHENGODE, NAMAKKAL TAMILNADU, INDIA
nithyameae@gmail.com

sumaiyakalamece@gmail.com

Abstract— Lung cancer is one of the most dangerous cancers in the world. These diseases can easily spread by uncontrolled cell growth in the tissues of the lung. Early detection of the cancer can save the life and survivability of the patients. There are various types of lung diseases such as Pneumonia, Tuberculosis, and Lung cancer, pulmonary fibrosis affecting chest wall, interstitium and alveoli. The lung disease has to be identified and classified correctly so that proper medical treatment can be provided. Various methods have been adopted in the detection of lung cancer. In order to detect the lung cancer in earlier stage, a computer aided diagnosis system is required. The aim of this work is to detect the lung cancer at earlier stage so that lung cancer patient's survival rate can be increased. The CT images of lung cancer patients are obtained before surgery. Lung cancer is the leading cause of cancer death in the World with survival restricted to a subset of patients even after undergoing surgical resection. The genomic biomarkers ERCC 1 [Excision Repair Cross Complementary 1] and RR [Ribonucleic Reluctase] were analyzed to identify the recurrence rate in lung cancer patients If recurrent, then patients are treated with Chemotherapy.

Index Terms— Computer-aided diagnosis, Fusion of image features and genomic biomarkers, Quantitative image feature analysis, lung cancer recurrence rate prediction

I.INTRODUCTION

LUNG CANCER is a serious health problem for our body. The mortality rate of lung cancer is highest among other types of cancer. It is one of the serious cancers with the smallest survival rate after the diagnosis. Survival from lung cancer is directly related to the growth at its detection time. The earlier the detection, higher the chances of successful treatment. In order to detect the lung cancer in earlier stage, a computer aided diagnosis system is required which can be implemented with image processing techniques.

From the recent years, medical CT Images have been used in medical field for diagnosis widely. In the recent years, medical CT Images have been applied in clinical diagnosis widely. It helps physicians to find and locate pathological changes with more precision. Computed tomography images can be distinguished for different tissues according to their different gray levels [1]. Lung diseases may be caused by infection, an outlet at the workplace, medications and various defects. X-ray chest radiography and computer tomography (CT) are two common anatomic imaging models that are mostly used in the detection and diagnosis of a various lung diseases. Genetic algorithms (GAs) are among the most famous methods to do a feature selection. .

CT images of the lungs produced are with noise and no obvious changes in the grayscale boundary and other features, and are not susceptible to split. Therefore, the first thing to do is preprocessing of an image. The aim is to remove these places which are not productive, which is to erase the noise of the lung CT images. Consequently, the boundary part of the CT image which changes smoothly should be sharpened in order to make possible the following work of segmentation [8]. The morphological smoothening and median filter techniques

are used to separate the noise from the images and improve the image. The median filter will detach the salt and the pepper noises and produces the improved image. In this work the darker details are overpower where the dilation is performed first and then it is followed by erosion.

The remainder of the paper is organized as follows: Section II briefly presents previous lung cancer identification methods. Section III describes the proposed framework. Section IV reports the experimental results and Section V offers the conclusion of the paper.

II.PREVIOUS WORK

A. LUNG NODULAR SEGMENTATION:

They used a semi-automated computer-aided detection (CAD) scheme to segment lung tumors. In brief, the tumor center and diameter that have been marked previously by a radiologist in the original clinical CT image reading were used as the initial segmentation seed in our CAD scheme. In each case, the image processing of the scheme started from segmenting the tumor area from the CT slice with the marks of the radiologist. The marked tumor seed was mapped into next adjunct CT image slices to segment the tumor area depicting on the next slices. This segmentation process was iteratively performed until the scheme reached a slice without remaining tumor area being detected. Specifically, in each involved image slice, CAD scheme first applied a conventional region growing algorithm with an empirically selected threshold of CT number (-450 HU). This step generally works well in segmentation of well-circumscribed lung nodules tumors, which connect with lung boundary and/or vessels. Thus, the following two image processing steps were applied. First, as shown in Fig. 1(a), by applying an initial region growing algorithm, the segmented tumor has a leakage to outside lung area (Fig. 1(b)). To segment this tumor, our scheme applied a modified convex hull function based algorithm introduced by Kuhgnik et al. [21] to stop the leakage of the segmented tumor area to the normal lung tissues and also smooth tumor boundary. Based on an anatomical fact that a lung is mostly convex, this algorithm is efficient and has good capability of removing the thoracic lesions from the chest wall (Fig. 1(c)). However, applying the convex hull function algorithm is also likely to generate a few minor (isolated) regions due to image noise. The scheme then applied a region labeling algorithm to remove small regions while maintaining the segmented tumor region (Fig. 1(d)).

Second, in order to cut and remove the connected vascular structures from the tumor boundary, our scheme applied a distance map based morphological operation as proposed by Kuhgnik et al. [21]. As shown , after a juxtapleural tumor (Fig. 1(a)) was segmented using the initial region growing algorithm and convex hull function, the scheme fitted a rectangular window to the initial tumor boundary (Fig. 1(b)). The window was also centered on tumor center. The scheme applied an Euclidean distance transform to convert the image inside the window into a distance map E that contains the minimum distance of each pixel of the tumor region to the tumor boundary pixels. Then, a seed optimization was done by searching for the pixel *C* with the longest distance in the neighborhood of the initial given seed.

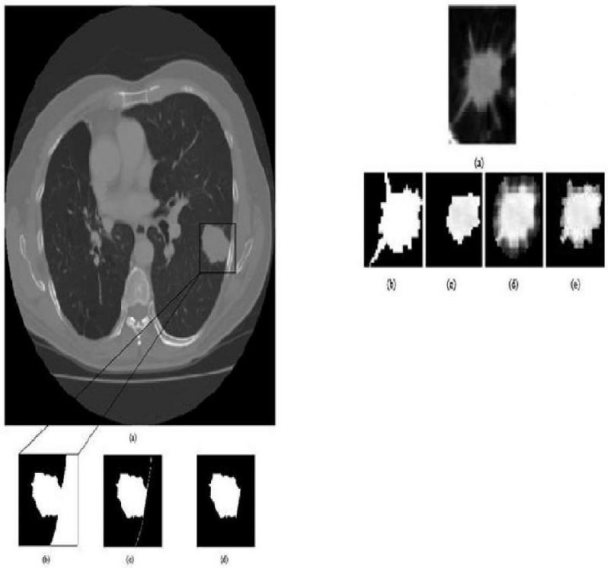

Fig.1. (a) A case with Juxtapleural tumor, (b) Leakage of region growing, (c) Chest wall connection is removed by convex hull function,
(d) Final segmented tumor region.

B. FEATURE EXTRACTION:

Image features Extraction stage is a crucial stage that uses algorithms and methods to detect and separate various preferred portions or shapes of an inputted image. The following two methods are used to predict the probability of lung cancer presence: binarization and PCM, both methods are based on facts that strongly related to lung anatomy and information of lung CT imaging.

(1) BINARIZATION APPROACH:

For detection of cancer binarization approach has been applied for detection of cancer. In binarization we extract the number of white pixels and check them against some threshold to check the normal and abnormal lung cells. After this process the condition is check whether number of white pixels of a new image is less than the threshold then it indicates that the image is normal, or else if the amount of the white pixels is greater than the threshold, it specifies that the image in abnormal. Merging Binarization and PCM methods together will lead us to take a decision whether the case is normal or abnormal.

(2) MASKING APPROACH:

Inside lungs masses are appeared as white connected areas inside ROI (lungs), masking approach depends on this. As they increase the percent of cancer presence increases. Also combining Binarization and Masking approaches together will help us to take a decision on whether the case is normal or abnormal according to the mentioned assumptions in the previous two approaches, we can make a conclusion that if image has number of black pixels greater than white pixels then that image is normal or otherwise we can say that the image is abnormal.

(3) PCA APPROACH:

PCA is a technique to normalize the data in image. Real-world data sets generally display associations among their variables. These associations are frequently linear, or at least practically so, making them agreeable to common examination techniques. One such technique is principal component analysis ("PCA"), which rotates the original data to new coordinates, making the data as "even" as possible. The features mined are delivered to the PCA data mining for better sorting [1].
The following steps takes place in PCA:-
i. Calculating the mean and standard deviation of the features in the image.

ii. Subtracting the sample mean from each observation, and then dividing by the sample standard deviation. This scales and centers the data.

iii. Then we calculate the coefficients of the principal components and their relevant changes are done by finding the Eigen function of the sample covariance matrix.

iv. This matrix holds the coefficients for the principal constituents. The diagonal elements store the modification of the relevant principal constituents. We can mine the diagonal.

v. The maximum variance in data results in maximum information content which is required for better classification [1].

(4) NEURAL NETWORK CLASSIFIER

Supervised feed-forward back-propagation neural network ensemble used as a classifier tool. Neural network contrasts

in different means from traditional classifiers like Bayesian and k – nearest neighbor classifiers. Linearity of data is one of the major variances. Other existing classifiers like Bayesian and k – nearest neighbor entails linear data to work properly. But neural network works as well for non linear data because it is simulated on the reflection of biological neurons and network of neurons. Training the neural network with wide range of input data will increase the detection accuracy, in other words the system will get biased with a small set of data or large set of similar data. Hence neural network classifier needs a large set of data for training and also it is time consuming to train to reach the stable state. But once it is trained it works as fast and quick as biological neural network by transmitting signals as fast as electrical signals.

Input layer, internal hidden layer and output layer are the three layers of the architecture of the neural network. The nodes in the input layer are linked with a number of nodes in the internal hidden layer. Each input node connected to each node in the internal hidden layer. The nodes in the internal hidden layer may connect to nodes in another internal hidden layer, or to an output layer. And the output layer consists of one or more response variables.
Following are the general Steps performed in Neural Network Classifier:-

i. Creating feed-forward back propagation network.

ii. Training neural network with the already available samples and the group defined for it.

iii. The input image mined PCA consistent data as the test samples, fires the neural network to check whether the particular selected input sample has cancer or not.

III. PROPOSED FRAMEWORK

A. GAUSSIAN FILTERING

Gaussian filtering is used to blur images and remove noise and detail.It is not particularly effective at removing salt and pepper noise. Gaussian filtering is more effective at smoothing images. It has its basis in the human visual perception system. It has been found that neurons create a similar filter when processing visual images. In one dimension, the Gaussian function is:

$$G(x) = (1/\sqrt{2}^{z})$$

Where z is the standard deviation of the distribution.The distribution is assumed to have a mean of 0. The Standard deviation of the Gaussian function plays an important role in its behaviour.

When working with images we need to use the two dimensional Gaussian function. This is simply the product of two 1D Gaussian functions (one for each direction) The Gaussian filter works by using the 2D distribution as a point-spread function. This is achieved by convolving the 2D Gaussian distribution function with the image. We need to produce a discrete approximation to the Gaussian function [13] . This theoretically requires an infinitely large convolution kernel, as the Gaussian distribution is non-zero everywhere. This means we can normally limit the kernel size to contain only values within three standard deviations of the mean. Once a suitable mask has been calculated, then the Gaussian smoothing can be performed using standard convolution.

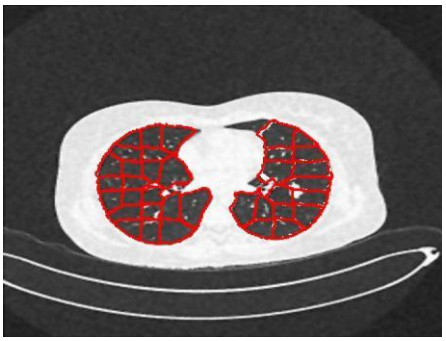

Fig 1.2 Filtered image

A pixel is considered to be a possible maximum if all neighboring pixels have lower or the same intensity, in which case pixels with the same intensity are stored in a queue, and tested in the same way. If eventually the queue is emptied so that all the pixels it contained proved to be possible maxima, then the corresponding connected component is a LMR. Pixels of a LMR are considered individually as possible candidates, and the [4] pixel with the maximum final score will represent the region; this procedure is referred to as non maximum suppression, and it will be discussed in details later in fig 1.2. We note that the usage of image smoothing, as discussed in the previous section, gains importance at this point, since the local intensity variations may be high on a raw retinal image, resulting in many local maxima.

B. GLCM (GREY LEVEL CO-OCCURRENCE METHOD)

The GLCM is a process of tabulating different combinations of pixel brightness values called as grey levels which occurs in an image. In this first step is to create gray-level co-occurrence matrix from image 1.3 in MATLAB.

$$P_{i,j} = \frac{V_{i,j}}{\sum_{i,j=0}^{N-1} V_{i,j}}$$

Where: i is the row number and J: is the column number. From this we compute texture events from the GLCM

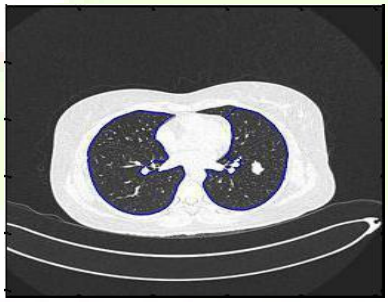

Fig 1.3 Brightened pixels

C. SUBSPACE BASED METHODS

Subspace based methods include the PCA, LDA, and ICA etc. The key idea behind PCA is to find an orthogonal subspace that preserves the maximum variance of the original data. [11] The PCA method tries to find the best set of projection directions in the sample space that will maximize the total scatter across all samples by using the following objective function:

$$J_{PCA} = \arg\max_{W} |W^{T} S_t W|,$$

where S_t is the total scatter matrix of the training samples, and W is the projection matrix whose columns are orthonormal vectors. PCA chooses the first few principal components and uses them to transform the samples in to a low-dimensional feature space.

LDA tries to find an optimal projection matrix W and transforms the original space to a lower-dimensional feature space. In the low dimensional space, LDA not only maximizes the Euclidean distance of samples from different

classes but also minimizes the distance of samples from the same classes. As a result, the goal of LDA is to maximize the ratio of the between-class distance against within [8] class distance which is defined as:

$$J_{LDA} = \arg\max_W (|W^T S_b W| / |W^T S_W W|)$$

where S_b is the between-class scatter matrix,

and S_W is the within-class scatter matrix. In the subspace palmprint identification method, the query palmprint image is usually classified into the class which produces the minimum Euclidean distance with the query sample in the low-dimensional feature space.D. ACTIVE SHAPE MODEL

Active shape models (ASM) are popular shape and appearance models for object localization. The method makes full use of prior of shape and appearance knowledge of object and has the ability to deform within some constraints. Based on modeling local features accurately, ASM obtains good results in shape localization .It depends mainly focus image description than intensity gradients, several image features are extracted to represent local image structure. One simple way to model features is to calculate the mean of each landmark in all training images. Then for every landmark, [1] the PCA is applied to model the variability of local features descriptors. When feature space is very high or the distribution of samples is non-Gaussian, the above modeling method doesn't work well. The K NN classifier with the selected set of features by sequential feature forward and backward selection is constructed at each landmark.

There are two main factors: To extract local edge orientation features using steerable filters as local image structure descriptions and To introduce a machine learning algorithm to model matching, which construct a classifier for each landmark by selecting a small number. In order to guide the matching between the model and the object, it is necessary to construct local appearance model for each landmark in the training examples .In ASM, the local image feature of each landmark is represented by sampling intensity gradients along the profile perpendicular to the landmark contour. It is assumed that these local image features are distributed as a multivariate [17] Gaussian for each landmark. Then the similar appearance model can be constructed by deriving the mean profile and the covariance matrix from the profile examples. The matching between the current positions in test image to the corresponding model is determined by minimizing the distance from the feature vector of the landmark to the corresponding model mean.

ASM is implemented by using the local appearance model to find a new shape and then updating the model parameters to best shape on each iteration. In order to obtain fast convergence, multi resolution framework is adopted. The model does match to the object in the way from the coarse to fine resolutions. The Active Shape Model (ASM) is a segmentation algorithm which uses a Statistical Shape Model (SSM) to constrain segmentations. This makes it possible to robustly segment organs with low contrast to adjacent structures. The standard SSM assumes that shapes are Gaussian distributed, which implies that unseen shapes can be expressed by linear combinations of the training shapes. Accurate segmentation of organs in medical images is challenging, because adjacent [5] structures are often mapped to the same range of intensity values, which makes it hard to detect their boundaries.

In these cases, prior knowledge of the shape of an organ can be used to avoid that the segmentation leaks into the neighboring structures. One of the most popular segmentation algorithms with a shape prior is the Active Shape Model (ASM), which uses a linear, landmark based Statistical Shape Model (SSM).The linear SSM is learned by a Principal Component Analysis (PCA) of the training shapes, which implies the assumption that the shapes are Gaussian distributed.

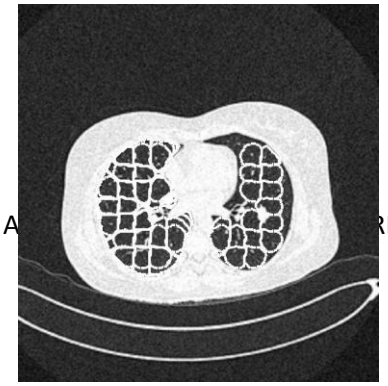

Fig 1.4 Robust ASM Matching

Image segmentation is the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics in the figure 1.4.

The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image. Each of the pixels in a region is similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristics.

When applied to a stack of images, typical in medical imaging, the resulting contours after image segmentation can be used to create 3D reconstructions with the help of interpolation algorithms like marching cubes.

E.SUPPORT VECTOR MACHINES

Support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

SVMs belong to a family of generalized linear classifiers and can be interpreted as an extension of the perceptron. They can also be considered a special case of Tikhonov regularization. A special property is that they simultaneously minimize the empirical *classification error* and maximize the *geometric margin*; hence they are also known as maximum margin classifiers.

A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

SVM is closely related to other fundamental classification algorithms such as regularized least-squares and logistic regression. The difference between the three lies in the choice of loss function: regularized least-squares amounts to empirical risk minimization with the square-loss,

$$\ell_{sq}(y, z) = (y - z)^2$$

logistic regression employs the log-loss,

$$\ell_{\log}(y, z) = \ln(1+e^{-yz})$$

The above features are doesn't dependent upon iteration method. So from this we can reduce computation time and also SVM here work as the Clustering based image retrieval so definitely our computation time must reduce.

IV. EXPERIMENTAL RESULTS

The study has a number of unique characteristics. First, although many different types of CAD schemes of chest CT images has been previously developed and tested [17], in this study we explored a new approach of applying a CAD-based quantitative image analysis concept to a new application field of predicting cancer recurrence risk. Since the ultimate goal of current promotion of lung cancer screening using low-dose CT imaging examinations is to reduce cancer mortality rate, this study investigated a clinically significant issue of how to more accurately predict cancer recurrence risk after initial surgery among the Stage I NSCLC patients. Our study results demonstrate the feasibility of developing a new CAD scheme to predict risk of cancer recurrence, which may eventually help improve the efficacy of lung cancer screening.

V. CONCLUSIONS

Developing precision medicine or a more effective personalized strategy for treating and managing Stage I NSCLC patients requires a more accurate clinical marker and/or assessment tool to predict cancer prognosis cancer recurrence risk). Current studies mainly focus on identifying more effective genomic biomarkers, demographic factors, and other clinical variables. In this study, we investigated a new quantitative image feature analysis approach using chest CT images and demonstrated two new study results. First, an image feature based classifier yielded significantly higher performance than two popular genomic biomarkers in predicting cancer recurrence risk. Second, the image features and genomic biomarkers are not highly correlated and provide supplementary information. As a result, fusion of these two types of features and biomarkers further improved prediction performance.

Despite the limitations, this preliminary study provides us a valid foundation to continue working in this new and promising CAD field to develop and optimize highly performed and robust risk prediction schemes that may have potential to eventually assist clinicians in more accurately identifying the patients with a higher risk of lung cancer recurrence after surgery. Therefore, for these high risk patients, the post-surgery chemotherapy is required to prevent or minimize the risk of cancer recurrence and thus increase their disease-free survival and overall survival time.

ACKNOWLEDGMENT

We would like to thank Mr. N.Senthil kumar for his valuable discussions and comments that improved the quality and clarity of the manuscript.

REFERENCES

[1] R. Siegel , *et al.*, ―Cancer statistics, 2015," *CA Cancer J Clin.,* vol. 65, pp.5-29, Jan/Feb., 2015.

[2] S. Swensen, *et al.*, ―CT screening for lung cancer: five-year prospective experience‖, *Radiology,* vol. 235, no. 235, pp. 259-265, APR. 2005.

[3] P. Bach, *et al.*, ―Computed tomography screening and lung cancer outcomes‖, *J. Amer. Med. Assoc.*, vol. 297, no. 9, pp. 953-961, Mar. 2007.

[4] L. Humphrey, *et al.*, ―Screening for lung cancer with low-dose computed tomography: a systematic review to update the U.S. Preventive Services Task

Force Recommendation‖, *Ann. Intern. Med.,* vol. 159, no. 6, pp. 411-420, Sep. 2013.

[5] S. Singhal., *et al.*, ―Prognostic implications of cell cycle, apoptosis, and angiogenesis biomarkers in non-small cell lung cancer: a review‖, *Clin. Cancer Res.*, vol. 11, pp. 3974-3986, 2005.

[6] R. Rosell, *et al.*, ―Screening for epidermal growth factor receptor mutations in lung cancer‖, *N. Engl. J. Med*, vol. 361, pp. 958-967, Sep. 2009.

[7] G. Bekler, *et al.*, "ERCC1 and RRM1 in the international adjuvant lung trial by automated quantitiative in situ analysis," *Am. J. Pathol.,* vol. 178, no. 1, pp. 69-78, Jan. 2011.

[8] J.C. Soria, ―ERCC1-tailored chemotherapy in lung cancer: the first prospective randomized trial‖, *J. Clin. Onc.*, vol. 25, pp. 2648-2649, 2007.

[9] G. Bepler, *et al*., "RRM1 and PTEN as prognostic parameters for overall and disease-free survival in patients with non-small-cell lung cancer," *J. Clin. Oncol.,* vol 22, no. 10, pp. 1878-1885, May. 2004.

[10] C. Poleri, *et al.*, "Risk of recurrence in patients with surgically resected stage I non-small cell lung carcinoma: histopathology and immune histochemical analysis," *Chest* vol. 123, no. 6, pp. 1858-1867, Jun. 2003.

[11] G. Loannidis, *et al*., "How close are we to customizing chemotheraphy in early non-small-cell lung cancer," *Ther. Adv. Med. Oncol.,* vol. 3, no. 4, pp. 185-205, Jul. 2011.

[12] E. P. Diamandis, "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations," *Mol Cell Proteomics,* vol. 3, no. 4, pp. 367-378, Apr. 2004.

[13] H. Li, *et al*., "Cost-effectiveness of a novel molecular test for cytologically indeterminate thyroid nodules," *J .Clin. Endocrinol Metab.,* vol. 96, no.11.

[14] S. Park, *et al.,* "Computer-aided detection of early interstitial lung diseases using low-dose CT images," *Phys. Med. Biol.,* vol. 56, no. 2, pp. 1139-1153, Jan. 2011.

[15] S. Park, *et al.,* "A multistage approach to improve performance of computer-aided detection of pulmonary embolisms depicted on CT images: Preliminary investigation," *IEEE Trans. Biomed. Eng.,* vol. 58, no. 6, pp. 1519-1527, June 2011.

[16] M. Tan, *et al.,* "A novel computer-aided lung nodule detection system for CT images," *Med. Phys.,* vol. 38, no. 10, pp. 5630-5645, Sep. 2011.

[17] D. Butcher, *et al*., ―A tense situation: forcing tumor progression,‖ *Nat. Rev. Cancer*, vol. 9, no. 2, pp. 108–122, Feb. 2009.

[18] F. Collins, H. Varmus, "A new iniative on precision medicine," *N. Engl. J. Med,* vol. 372, no. 9, pp. 793-795, Feb.

Ms.M.Nithya M.E. Assistant professor ECE had presented a paper ―LUNG CANCER RECURRENCE RATE BY GENOMIC BIOMARKERS‖ on conference held on KIOT .and published the same on R.K..Publications. An ISTE Student member.

K.Sumaiya parveen pursuing M.E. Embedded system Technologies had presented a paper on ―Digital image processing in todays scenario‖ and published journal on ―LUNG CANCER RECURRENCE RATE BY GENOMIC BIOMARKERS‖ on R.K..Publications. An ISTE Student member