A Comparative Study for Comparing Two Feature Extraction Methods and Two Classifiers in Classification of Early-stage Lung Cancer Diagnosis of chest x-ray images

Amal M. Al Gindi [1,*], Tawfik A. Attiatalla [1] and Mostafa-Sami M. Mostafa [2]

[1] Department of Mathematics, Faculty of Science, Suez Canal University, Ismailia 41522, Egypt
[2] Department of Computer Science, Faculty of Computers and Informatics, Helwan University, Cairo, Egypt
algindi_a@yahoo.com

Abstract: ***Background:*** Lung cancer is one of the most common cancer-related deaths. Although attention has been paid to early stage predictions and diagnoses, prognosis remains very poor. This problem can be approached by developing more discriminative diagnosis methods. **Purpose:** In this paper a computer-aided diagnosis is proposed to solve the problem of classification of solitary pulmonary nodules in chest x-ray images for diagnosis of early-stage lung cancer in x-ray lung images. ***Methods:*** A set of 247 chest x-ray images from Standard Public Database by Japanese Society of Radiological Technology (JSRT) database used with 93 non-nodule images, 100 malignant and 54 benign images. Curvelet transform has been used in the process of feature generation and extraction and compared with that of Wavelet transform used for the same purpose in a previous research of our group in case of Euclidean distance classifier. A Support Vector Machine-based classifier prediction model is established and compared with Euclidean distance classifier that used for the same purpose based on Curvelet features. Validation of the classification is performed using a HoldOut method, while evaluation of the classification performance is computed and compared with other research's results in this area. ***Results:*** Using Curvelet transform for the process of feature generation and extraction, support vector machine prediction model is more effective for lung cancer detection since it increases the rate of diagnosis for early-stage lung cancer in x-ray lung images. ***Conclusions:*** Using Support vector machine in the process of diagnosis of solitary pulmonary lung nodules is more sensitive for diagnosis of early-stage lung cancer in chest x-ray images than Euclidean distance classifier when the feature extraction is based on Curvelet than that wavelet-based models.

[Amal M. Al Gindi, Tawfik A. Attiatalla and Mostafa-Sami M. Mostafa **A Comparative Study for Comparing Two Feature Extraction Methods and Two Classifiers in Classification of Early-stage Lung Cancer Diagnosis of chest x-ray images.** *J Am Sci* 2014; 10(6):13-22]. (ISSN: 1545-1003). http://www.jofamericanscience.org. 3

Keywords: x-ray images, lung nodules, feature extraction, Wavelet transform, Curvelet transform, Euclidean distance, Support Vector Machine.

I. Introduction

Lung cancer is one of the most harmful forms of cancer, which is the leading cause of cancer death in many regions of the world [1]. Early detection of lung cancer is essential in reducing life fatalities. However, achieving this early detection of lung cancer is not an easy task. More than 80% patients are already in middle or advanced stage when diagnosed and they miss the timing for the surgery. The 5-year survival rate is only 14%, which can reach more than 70% if lung cancer can be diagnosed in an earlier stage [2]. This difficulty in diagnosis at the early period explains the need for an early stage prediction model.

Interpreting a chest radiograph is extremely challenging. Superimposed anatomical structures make the image complicated, so even experienced radiologists have trouble distinguishing infiltrates from the normal pattern of branching blood vessels in the lung fields, or detecting subtle nodules that indicate lung cancer [3]. Chest radiography is the most frequently used diagnostic imaging examination for chest diseases such as lung cancer, tuberculosis, pneumonia, pneumoconiosis, and pulmonary emphysema. More than 9 million people worldwide die annually from chest diseases. Lung cancer causes 945000 deaths, and is the leading cause of cancer deaths in the world and in countries such as United States, the United Kingdom, the Russian Federation, Canada, Poland and Japan [4] .

Early detection is the most promising strategy to enhance a patient’s chance of survival. Early detection can be achieved in a population screening, the most common screening for lung cancer make use of chest projection radiography, or low-radiation dose Computer Tomography (CT) scans. It has been shown in the Early Lung Cancer Action Project that low-dose CT is more effective than conventional chest X-ray for the detection of pulmonary nodules. Moreover, it was found in a lung cancer screening for heavy smokers, that when radiographs were checked in retrospect, 90% of peripheral lung cancers nodules were visible. In fact lung cancer missed on chest radiographs is the second most common reason for litigation against radiologists.

Although histology diagnosis is the most accurate detection method in the medical environment, it is an aggressive invasive procedure that involves some risks, patient discomfort and some trauma, which restricts it to be used in the clinical practice. Digital CT, overcoming shortages of histology diagnosis, has gradually become the best imaging diagnosis method of lung cancer. But pulmonary nodules (referring to the lesion of lung field ≤ 3 cm in diameter) of lung cancer in CT images share similarity with benign cases to some extent, such as tuberculosis, inflammatory pseudotumor, hamartoma, and aspergillosis, which makes it difficult to distinguish, especially for the doctors who are not rich in clinical experience [2].

Although CT scans is more effective than conventional chest X-ray for the detection of pulmonary nodules as mentioned before chest radiographs (CXRs) are used far more commonly for chest diseases because they are the most cost-effective, the most routinely available, and the most dose-effective diagnostic tool, and they are able to reveal some unsuspected pathologic alterations [6]. Because CXRs are so widely used, improvements in the detection of lung nodules in CXRs could have a significant impact on early detection of lung cancer. Although CXRs are the most widely used modality for lung diseases, it has been well demonstrated that detection of lung cancer at an early stage in CXRs is a very difficult task for radiologists. The difficulties in detecting lung nodules in CXRs are threefold: (1) There is a wide range of nodule sizes, (2) nodules exhibit a large variation in density in CXR, and (3) nodules can be obscured by other anatomic structures. The reasons for misdetection may be due to difference in decision techniques, lack of clinical data, and structured noise in CXRs [7].

For these reasons there is a particular interest for the development of computer algorithms that can serve as a second reader, highlighting suspicious regions in the radiographs that then have to be judged by a radiologist [5]. The computer-aided diagnosis (CAD) has become an auxiliary diagnosis tool, especially in diseases that cannot be diagnosed efficiently [2]. Computer aided detection of solitary pulmonary nodules is faced with many difficulties due to the existence of complicated anatomical structures in chest radiographs. Automatic classification of the regions of interest (ROI) as nodules needs to extract a class of powerful region-based features. The performance of region-based feature extraction hinges on a successful nodule segmentation algorithm of suspicious regions [8].

To improve the accuracy and efficiency of CT screening programs for the detection of early-stage lung cancer, a number of research projects, such as texture analysis and image segmentation, have been done to assist radiologists in diagnosing lung cancer. The purpose of our research is first, to compute the classification performance of using Curvelet as feature extraction tool with Euclidean distance classifier and compare the results with that which demonstrated by our group in [15] in which we used Wavelets for feature generation and extraction process and the same classifier. Second, to establish a Curvelet-based algorithm to extract texture features of X-ray images and compare the diagnosis rate in case of two classifiers Euclidean distance and Support Vector Machine to decide which one is more effective to be the prediction model for diagnosis of early-stage lung cancer.

2. Material and Methods

A. Material

A Standard Public Database by Japanese Society of Radiological Technology (JSRT) database which is publicly available have been used in applying the new scheme [9]. This database is selected according to the variety cases included and widely usage in similar research work. A set of 247 chest x-ray images which is original posteroanterior chest films (34.6 cm × 34.6 cm) for this database were collected from 13 medical centers in Japan and one institution in the United States as follows: one nodule per image for nodule cases, all of the original radiographs were digitized using an LD-4500 or an LD-5500 laser film digitizer. Digitized images had a 2048 × 2048 matrix, 0.175-mm pixel size, and 12-bit gray levels. The database included 247 posteroanterior chest images, which consisted of 154 images with and 93 images without a nodule. One hundred nodules were malignant and 54 were benign, All images were presented in a randomized sequence for detection of lung nodules. Figure 1 shows the distribution of nodule sizes in JSRT and table 1 shows the Gender Distribution of JSRT nodule images.

B. Methods

We computed the Curvelet-based proposed scheme using a program code written in Matlab® version 7.9.0.529 (R2009b) with the use of Matlab® Image processing toolbox and using Curvelet toolbox as a multiscale level of decomposition to represent pulmonary nodules of x-ray images. Also, we used and Microsoft Visual C++ software for conversion of database images 16-bit (BIG endian) to 32-bit (Little endian) format for much better image enhancement.

In this article, Curvelet texture analysis was used with Support Vector Machine to establish a prediction model for detecting early-stage lung cancer in chest x-ray in addition to a comparative study for a two texture feature extraction tools and two classifiers, which has not been reported to our knowledge since nearly all the recent studies are dealing with CT images.

This article is organized as follows, section 1 is an introduction to early detection of lung cancer and difficulties facing radiologists especially in x-ray images. Section 2 the materials and methods are proposed, which includes the database used in training and testing the system and some of its details, the software and programming languages used for implementing the system and some of the used concepts . Section 3 presents a brief description of classification methods, an evaluation of CAD schemes and the proposed scheme. The experimental results achieved and its discussion is presented in section 4 while conclusion is presented in section 5.

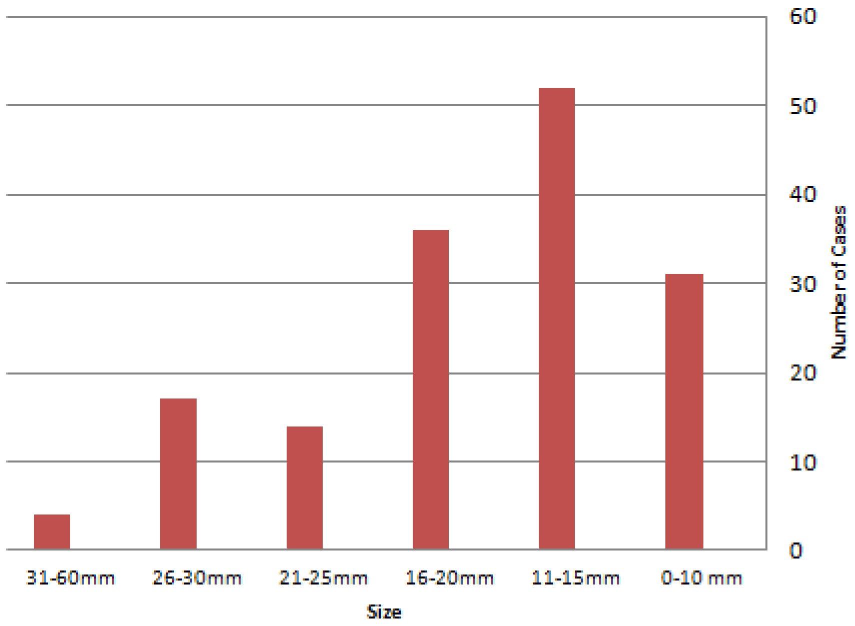

Fig. 1 The number of cases and the corresponding nodule sizes in JSRT database

Table 1 Gender Distribution of JSRT Nodule Images

	Number of	Nodule	Gender		Total
	Cases	Type	Male	Female	
Nodule Cases	154	Benign	27	27	54
		Malignant	41	59	100
Total			68	86	154

C. Main Concepts

1. Feature Extraction

Dimensionality reduction is the process of reducing the number of random variables (features) under consideration, and can be divided into feature selection and feature extraction. In image pattern recognition, feature extraction is the first step in image classification; it is a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed then the input data will be transformed into a reduced representation set of features (features vector). Transforming the input data into the set of features is called feature extraction. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. In [10], a total of 114 features for each nodule candidate including geometric, intensity and gradient features were computed. An inside-outside-feature separation process was applied, the subset of features is chosen using a sequential forward selection (SFS) process, which based on the area under the free receiver operating characteristic (FROC) curve for the chosen classifier. The results indicate that the system is able to detect 78.1% of the nodules in the JSRT test set. Another approach for reducing dimensionality in CT scan images was proposed in [11], showed that since features in different dimensions might provide useful information for the characterization, a system of three dimensions of image features (2D, 2.5D and 3D features) had been used. Features for grayscale, shape, invariant moment, gradient, and texture features were calculated for the nodules and the surrounding areas in 2D, 2.5D and 3D dimensions. Extra features such as histogram of grayscale features, texture features and gradient features were added.

2. Texture Extraction

Texture is a fundamental characteristic of the digital images as it usually reflects the structure of the pictured objects. The methods of texture extraction can be classified into four parts: statistical method, model method, spectrum method and structural method. The basic procedure of texture analysis is to extract texture of images using different methods and then run a set of mathematical texture operators to produce a corresponding set of texture feature values in order to describe character of images [1]. The Wavelet transformation, a textural features extraction method, provides a multi-resolution and non-redundant representation of signals with an exact reconstruction capability, and forms a precise and uniform framework for the space–frequency analysis [12].

2.1 Curvelet Transform

Although Wavelets perform very well for objects with point singularities, they are not adequate for representing 1D singularity [12]. The success of wavelets lies in its good performance for piecewise smooth functions in one dimension, however wavelet is not suitable to capture more directional features in an image but since 2D images are irregular when decomposed, Curvelet transform is more suitable than the wavelet transform to extract texture features [2]. In 2000 Curvelet was developed, a type of second generation Wavelets. As an extension of the Wavelet multiscale analysis framework, Curvelets can effectively deal with linear singularities in 2D signals. The Curvelet transformation is defined as an effective tool for finding curves at multiple resolution levels. Several studies using Curvelet transformations in image processing have shown that Curvelet transformations yield better results [12]. Curvelet transform, is a kind of spectrum method, stems from Wavelets theory, but it overcomes the weakness of traditional multiscale representations using wavelets, and is suitable to capture more directional features in an image. The main formulas offering to Curvelet transform are as follows [1]

$$\emptyset_{j,l,k}\,(X) = \emptyset_j(R_{\theta_1}(X - X_k^{(j,l)}\,))$$

where R_θ is the rotation by θ radians and R_θ^{-1} its inverse

$$R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, R_\theta^{-1} = R_\theta^T = R_{-\theta}$$

A Curvelet coefficient is the inner product between an element $f \in L^2\,(R^2\,)$ and a Curvelet $\theta_{\,j\,,l\,,k}$

$$C(j,l,k)" := \int_{R^2} f(X)\,\emptyset_{j,l,k}\,(X)dx$$

Where R denotes the real line.

The basic idea of Curvelet transform is to represent a curve as a superposition of functions of various lengths and widths obeying a specific scaling law. Regarding 2D images, it can be done first by decomposing an image into wavelet sub-bands, i.e., separating the object into a series of disjoint scales. Each sub-image of a given scale is then analyzed with a local ridgelet transform, another kind of new multi-resolution analysis tool [2]. Discrete Curvelet Transform (DCT) is a new image representation approach that codes image edges more efficiently than wavelet transform. Indeed, Curvelet has useful geometric features that set them apart from wavelet [13].The Curvelet transform coefficients of the object are used as a feature vector. Suppose we have a function f which has a discontinuity across a curve, and which is smooth otherwise, and consider approximating f from the best m-terms in the expansion. The squared error of such an m-term expansion is given by:

$$\|f - f_{\tilde{F}}\|^2 \quad \alpha\ \frac{1}{\sqrt{m}}, \quad \text{m} \rightarrow +\infty$$

($f_{\tilde{F}}$ is the approximation from m best Fourier coefficients). This shows that the mean squared error will be reduced in the Curvelet. Figure 2 shows the Curvelet tiling in the frequency domain.

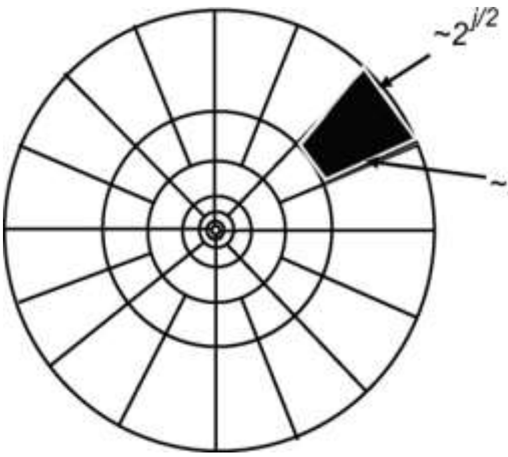

Fig. 2 Curvelet tiling in the frequency domain

In [13] a promising use of Curvelet transform with mammogram images. It had been used at different scales as a preprocess for feature extraction and classification of mammogram images, then extracting different ratios of the biggest Curvelet coefficients from each level as a feature vector to be used for classification. Euclidean distance based classifier used for classification process. In the proposed scheme we used the same algorithm in this article by selecting not all the largest coefficients but only an amount of 10% of them and developed it also by setting the other coefficients to zero after sorting them decendingly. We applied a similar technique before in a previous article [15] using 4-level Wavelet transform for feature extraction process and computing the mean of the resulting coefficients at each level then the average instead of computing the biggest coefficients. Here we compute the system performance when using Curvelet transform as feature extraction tool in two cases first, with Euclidean distance classifier and compare it with the results we obtained in [15]. Second, with Support Vector Machine classifier and then compare it with that of Euclidean distance using the extracted texture features in both cases with different percentages to establish the SVM prediction model.

III. CLASSIFICATION

The classification step is the final step in our model where both the features of the training images and the test images are the input of the classifier, while the output is the image type. In [14] features used for classification are taken from a Multiscale Gaussian filterbank. The complete set of features consists of a total of 109 features. In their multiscale CAD scheme for detecting pulmonary nodules in chest radiographs they found that the inclusion of the candidate selection step had a clear positive effect on system performance. The effect of the candidate segmentation step was less apparent. The methods of establishing prediction model are variable, such as logistic regression, discriminant analysis, artificial neural networks, and support vector machine [1], next are little details of some of them,

1. Minimum Distance Classifier

The minimum distance classifier is used to classify unknown image data to classes which minimizes the distance between the image data and the class in multi-feature space. The distance is defined as an index of similarity so that the minimum distance is identical to the maximum similarity, the following distances are often used in this procedure,

1.1 Euclidian distance

This is given by the formula

$$d_k^2 = (\mathrm{X} - \mu_k)^{\mathrm{t}} \cdot (\mathrm{X} - \mu_k)$$

This is used in cases where the variances of the population classes are different to each other, i.e. it acts as similarity index.

In this study we design an Euclidean distance-based classifier as a first phase of classification or as a first classifier, similar to that proposed in [15], that is based on calculating the distance between the feature vectors for the input testing images and the class core vector obtained from the trained set of images using 25% of the total number of coefficients resulted by applying two scales curvelets decompositions. The system automatically classifies the feature vector to a diagnosis class $\mathrm{C}_{\mathrm{diag}}$ by finding the nearest class to this vector. This is done by testing the distance between this feature vector and all class core vectors, the following equations were used to the classification process

$$\mathrm{Dist}(A, C_{diag}) = \mathrm{MinDist}, \quad (1)$$

$$\mathrm{MinDist} = \min_{1 \ll \mathrm{m} \ll \mathrm{M}}\left(\mathrm{Dist}(A, C_m)\right), \quad (2)$$

$$Dist(A, C_m) = \frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{L^j}\sqrt{(A^j\,(i) - C_m^j\,(i)^2}\,, \quad 1 \ll m \ll M \quad (3)$$

While A_j is the coefficient vector of the jth decomposition level for diagnosis image, C_m^j is the class core vector for class m at decomposition level j, L^j is the length of coefficient vector at decomposition level j, M is the number of classifications classes, and J is the number of decomposition levels used. The class core vector for each experiment were calculated using the following equation,

$$C_m^j = \frac{1}{N^j}\sum_{n=1}^{N^j}\sum_{i=1}^{L^j} A_m^j\,(i), \quad 1 \ll m \ll M, \quad (4)$$

Where N_j is the number of selected ROI's to produce the class core vector at decomposition level j, and A_m^j is the coefficient vector for ROI's for the class m at decomposition level j.

2. Support Vector Machine

The final phase in the proposed algorithm is the classification of occurrence and non-occurrence of cancer nodule for database lung images using Support Vector Machine (SVM). SVM is a modern outgrowth of artificial neural networks. SVM model using a sigmoid kernel function is equivalent to a two-layer, feed-forward neural network. SVM is usually used for classification tasks, in case of binary classification SVM is used to find an Optimal Separating Hyper plane (OSH) which generates a maximum margin between two categories of data. To construct an OSH, SVM maps data into a higher dimensional feature space.

SVM performs this nonlinear mapping by using a kernel function [16].

Support vector machine is based on the structural risk minimization principle [17]. SVM approach enjoys many attributes, it is less computationally intense in comparison to artificial neural networks. It performs well in high-dimensional spaces and also well on both training data and testing data but does not suffer from the small size of training dataset as do other kinds of classifiers since the decision surface of SVM-based classifier is determined by the inner product of training data. The basic idea of SVM is to construct a hyperplane that maximizes the margin between negative and positive examples. The hyperplane is determined by the examples called support vectors that are closest to the decision surface. The decision surface is determined by the inner product of training data, which enables us to map the input vectors through function Φ into a higher-dimensional inner product space called feature space. The feature space could be implicitly defined by kernel K(x, y). To tolerate noise and outliers and to avoid overfitting, slack variables ξ_i are introduced which allows the margin constraints to be violated.

Consider the training samples (x_i, y_i), i=1,…,m, where each point x_i is an input vector with label $y_i \in \{-1, 1\}$. The decision surface has the form:

$$y = k(x, w) + b$$

The decision surface is the solution of the following optimization problem:

$$\text{minimize: } \frac{1}{2}\, k(w, w) + C \sum_{i-1}^{l} \mathcal{E}i$$

$$\text{subject to : yi } [k(w, x_i)+b] \geq 1 - \mathcal{E}_i,\ i = 1\ldots l$$

$$\mathcal{E}_i \geq 0,\ i = 1\ \ldots l$$

Where C > 0 is a parameter chosen by the user for decision errors.

3. Performance Evaluation

In classifier construction studies for lung CAD, the ANN was usually used, so a comparison for the performance of SVM-based classifier with ANN-based has been carried out in [17]. Employing a two-layered feed forward neural network that contains one input layer, one hidden layer, and one output layer, comparing the performance of SVM and back propagation (BP)-ANN in differentiating solitary pulmonary nodules using the selected feature subsets with leave-one-out procedure, they found that the performance of SVM-based classifier in differentiating SPNs is better than that of the ANN-based classifier. A novel fast marching approach was developed in [18], which combines the prior knowledge of pulmonary nodule boundary into the nodule segmentation algorithm via the SVM classifier, to calculate decision value of each pixel in suspicious region in chest radiograph. The proposed watershed segmentation algorithm implements an adaptive velocity function according to the local image gradient and intensity, which is efficient for SPNs with weak boundaries. In [19], a nonlinear SVM with a Gaussian kernel was employed for classification of the nodule candidates. Using Gaussian kernel among several kernels is because it achieved the best performance. The SVM was trained/tested with a leave-one-out cross-validation test. Comparing the nonlinear SVM classifier with LDA (Linear Discriminant Classifier), the SVM generalized from a relatively small number of positive cases did better than did the LDA.

In [1] a support vector machine prediction model was established for small pulmonary nodules using Curvelet transform to extract texture features of CT image using two examples, example 1: was multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image using gray level co-occurrence matrix to get fourteen textural features. Example 2: Support vector machine prediction model for small pulmonary nodules based on Curvelet transform to extract texture features of CT image in order to promote the ratio of detection and diagnosis of early-stage lung cancer. Results show that the classification consistency, sensitivity and specificity for the model are 81.5%, 93.8% and 38.0% respectively. A CAD scheme was proposed in [12] for Early-Stage Lung Cancer. The synthetic minority over-sampling technique (SMOTE) was used for raw data in order to balance the original training data set. Curvelet-transformation textural features, together with 3 patient demographic characteristics, and 9 morphological features were used to establish a support vector machine (SVM) prediction model. Longitudinal data as the test data set was used to evaluate the classification performance of predicting early-stage lung cancer. Accuracy based on cross-evaluation for the original unbalanced data and balanced data was 80% and 97%, respectively was reached.

In the proposed algorithm validation of the classification is performed using a HoldOut cross validation method, while evaluation of the classification performance is based on comparing the results with other research's results in this area.

4. The Proposed Scheme

Figure 3.a shows the main steps for the first stage of the comparative system, while figure 3.b shows the main steps of the second comparative system. The proposed CAD scheme for lung nodule classification in x-ray images consists of (1) a two-stage image enhancement technique (2) feature generation using Curvelet transform (3) nodule selection and extraction (4) feature extraction of the candidate ROI's using ROI size 128 x 128 pixels, (5) dimensionality reduction of the selected features through selecting the biggest coefficients, sorting them and setting the other coefficients to zero (6) classification of the nodule candidates into malignant or benign using two classifiers Euclidean distance and Support Vector Machine (7) comparing the two feature extraction tools in case of Euclidean distance (8) comparing the two classifiers in case of Curvelet texture extraction tool.

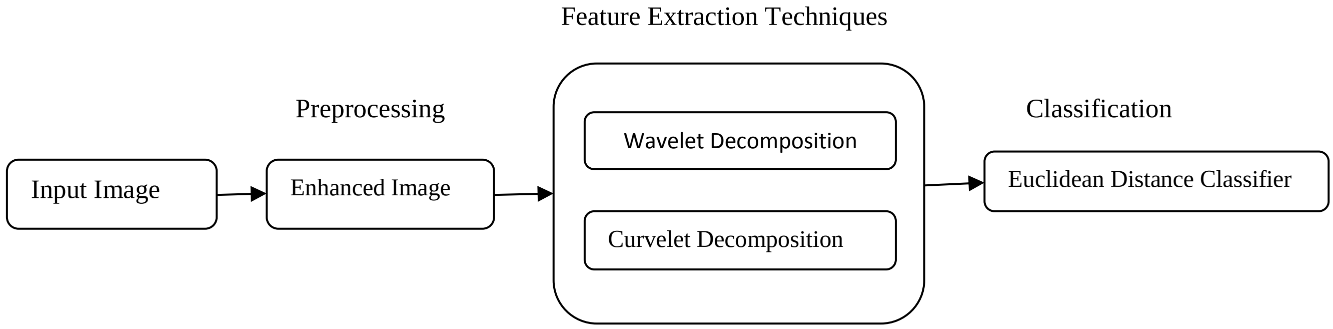

Fig. 3. a The first comparative system using 2 feature extraction methods and 1 classifier

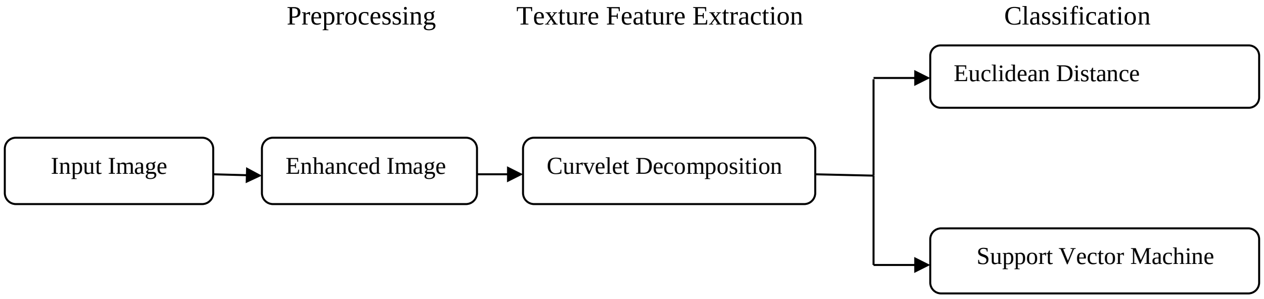

Fig. 3. b The second comparative system using Curvelet feature extraction and 2 classifiers

4. Preprocessing

4.1 High-Frequency Emphasis filtering (HFEF)

In which an approach to compensate for the problem of losting background tonality of the 32-bit (Little endian) format converted image, is to add an offset to a highpass filter. When an offset is combined with multiplying the filter by a constant greater than 1, the approach is called *High-Frequency Emphasis* filtering because the constant multiplier highlights the high frequencies. The multiplier increases the amplitude of the low frequencies also, but the low-frequency effects on enhancement are less than those due to high frequencies .as long as the offset is small compared to the multiplier. High-frequency emphasis has the transfer function

$$H_{hfe}(u, v) = a + bH_{hp}(u, v)$$

Where a is the offset, b is the multiplier, and $H_{hp}(u, v)$ is the transfer function of a highpass filter. In a previous study, the same technique was demonstrated by our group [15]. A Butterworth highpass filter of order 2 and standard deviation value equal to 5% of the vertical dimension of the padded image width were used, then using high-emphasis filtering with a = 0.5 and b = 2.0 we got an advantage for the resulted image in which the gray-level tonality was retained, and we apply a histogram equalization for the resulted image for further enhancement.

Figure 4 show the difference between the image before and after applying the high-frequency emphasis filter and histogram equalization.

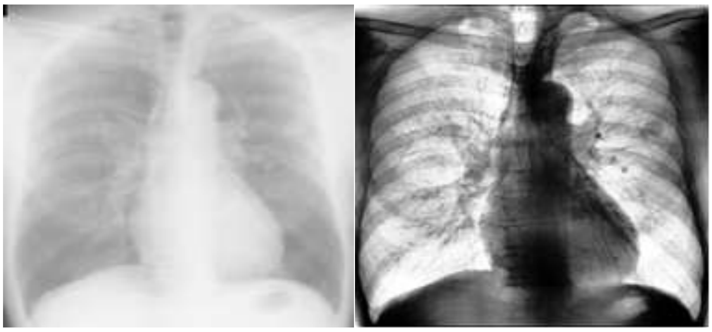

Fig. 4 Enhanced lung field by using high frequency emphasis filtering

The result obtained using a combination of high frequency emphasis and histogram equalization is superior to the result that would be obtained by using either method alone. From the selected images in the JSRT database, the following steps were carried out,

4.2 Nodule Selection and Extraction

(i) Nodule Selection, a regions of interest (ROI's) were first selected depending on that the locations of the nodules were confirmed by three chest radiologists who were in complete agreement as we mentioned above.

(ii) Nodule Extraction, ROI's were extracted with size128 x 128 pixels, these selected nodule size is because the average size of all nodules included in the database was 17.3 mm. and the average pixel size is 0.175- mm, i.e. 128 x 128 pixels nearly equal 22.4 mm, which permits to analyze nodules with some surrounding areas . The reason for selecting nodules with some surrounding areas is that radiologist' observations on the benign and malignant nodules and their background suggests that the surrounding areas may carry information which might differentiate them. [15]

IV. Results

As a pre-process for classification, a Curvelet transformation produced a multi-scale level of decomposition for 50% of x-ray database images. The two and three scales' Curvelet coefficients matrices (the coarse layer, the detail layer, and the fine layer) were chosen as candidates and since we found no significant difference in classification results between scale 2 and scale 3 we considered the results of scale 2 only in our results. ROI images were decomposed, resulting in the extraction of a maximum number of 32768 of textural features from each ROI. A percentage of 10% of the biggest coefficients were used in the training phase and setting the others to zero and a percentage of 25% nearly 8192 feature of the total number of features is used in the testing phase, since the reasonable method to compare feature extraction methods to evaluate their performance is to use the same classification system, and the same database images. We used 50% of the database images randomly as a training set and 30% as testing set used to evaluate the validity of our prediction model. Classification performance was performed in order to compare the classification results for the Curvelet-based method to that of the wavelet-based method and to those that reported by other researchers.

Table 2 shows that using Euclidean distance classifier we got better results with 25% of Curvelet decomposition texture features than that of using 100% of Wavelet decomposition (bior1.5) features as we obtained a correct classification rate of 97% and 95% respectively. Performance comparison of the two considered classifiers is introduced in table 3. It is clear that the better results achieved using Euclidean distance classifier with only 25% of Curvelet texture features and it was 97% as mentioned before while using the whole number of features i.e. 100% of them using SVM classifier we achieved an average result of 98.46 for seven randomly cases of different numbers of testing images from the database shown in table 4. In the first case we used the whole database i.e. 100 malignant & 54 benign images which result in correct rate and sensitivity of 100% and 100% respectively. And since the smaller number of benign cases makes the training sample lack of benign data we used in the rest of cases an equal number of malignant and benign cases, so using 100 images as 50 malignant and 50 benign resulting in 98% and 96% for correct rate and sensitivity respectively. Then using 90 images with equal number of cases we got the same results as using the whole database i.e. 100% and 100% correct rate and sensitivity respectively. Also, it is clear that using 80 images of both cases we achieved 97.5% and 95% respectively, and using 70 images results in 97.06% and 94%. Also, using 60 equally divided cases images we obtained 96.67% and 93% respectively. Finally using 50 images which is corresponding to about 32% of the whole database as testing set the results was 100% and 100% correct rate and sensitivity respectively, i.e. results showed that Curvelet decomposition and SVM showed the best classification performance.

The high sensitivity (true positive rate) obtained, which is always the goal of this kind of work can reduce the false negative rate of early-stage lung cancer effectively.

Table 2. Correct classification rate, using Euclidean distance classifier and the number of coefficients in each experiment in percentage

Classifier	Feature Extraction Method	No. of Coefficients (%)	Correct Rate (%)
Euclidean distance	Wavelet Decomposition (bior1.5)	100%	95%
	Curvelet Decomposition	25%	97%

Table 3. Correct classification rate, using Curvelet Decomposition and the number of coefficients used for each classifier in percentage

Feature Extraction Method	Classifier	No. of Coefficients (%)	Training Set (%)	Testing Set (%)	Correct Rate (%)
Curvelet Decomposition	Euclidean distance	25%	50%	30%	97%
	SVM	100%	50%	50%	100%

Table 4. Correct classification rate for linear Support Vector Machine classifier using different number of images through the database

No. of images	Correct Rate (%)	Sensitivity
154 (All the database images i.e. 100 malignant & 54 benign)	100%	100%
100 (50 benign & 50 malignant)	98%	96%
90 (45 benign & 45 malignant)	100%	100%
80 (40 benign & 40 malignant)	97.5%	95%
70 (35 benign & 35 malignant)	97.06%	94%
60 (30 benign & 30 malignant)	96.67%	93%
50 (25 benign & 25 malignant)	100%	100%

V. Conclusion

In this article, a novel model for automatic classification of pulmonary lung nodules in x-ray images is proposed. In order to select texture features which are more accurate to reflect characteristics of pulmonary nodules, we have made two attempts. The first one was proved in a previous work of our group using Wavelet decomposition and the second using Curvelets decomposition for multiresolution analysis of lung nodules then comparing the results in case of using Euclidean distance classifier. A new idea for x-ray lung nodule images is explored using another classifier which is Support Vector Machine classifier and comparing the classification performance using the two classifiers based on Curvelet texture features. Experiment is applied on real labeled data. Since it is challenging to distinguish between benign and malignant cases especially in x-ray images, and it increases in case of lack of experience, it was the need for CAD system which promote the classification accuracy, in this article we developed such CAD system and results show promising use of this technique. The performances of the classifiers in terms of sensitivity and classification accuracy are shown. The results indicated that the SVM approach yielded the better performance when compared to the Euclidean distance classifier based on Curvelet texture feature than that based on Wavelet texture feature.

References

[1] Xiuhua G., S. Tao, W. huan and L. Zhigang, " Theory and Applications of CT Imaging and Analysis- Prediction Models for Malignant Pulmonary Nodules Based-on Texture Features of CT Image," InTech, 63-76 , (2011).

[2] Xiuhua G., S. Tao, W. Haifeng, *et al.*, " Support Vector Machine Prediction Model of Early-stage Lung Cancer Based on Curvelet Transform to Extract Texture Features of CT Image," World Academy of Science, Engineering and Technology 47, 333-337, (2010).

[3] Ginneken B. V., B. M. H. Romeny and M. A. Viergever," “Computer-Aided Diagnosis in Chest Radiography: A Survey," IEEE transactions on medical imaging, vol. 20, no. 12, pp. 1228-1241, 2001.

[4] Suzuki K., H. Abe, H. MacMahon and K. Doi, "Image Processing Technique for Suppressing Ribs in Chest Radiographs by Means of Massive Training Artificial Neural Networks (MTANN)," IEEE transactions on medical imaging, vol. 25, no. 4, pp. 406-416, 2006.

[5] Schilham A. M. R., B. V. Ginneken and M. Loog, "A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database," Medical Image Analysis, vol. 10, pp. 247-258, 2006.

[6] Murphy G. P., Lawrence W., Lenhard R. E., *American Cancer Society Textbook of Clinical Oncology*, 2nd edition, The Society, Atlanta, GA, 1995.

[7] Chen S., K. Suzuki and H. MacMahon, "Development and evaluation of a computer-aided diagnostic scheme for lung nodule detection in chest radiographs by means of two-stage nodule enhancement with support vector classification," Med. Phys. Vol. 38, no.4, pp. 1844-1858, 2011.

[8] Coppini G., S. Diciotti, M. Falchini, N. Villari, G. Valli, "Neural networks for computer-aided diagnosis: detection of lung nodules in chest radiograms,"*IEEE Transactions on Information Technology in Biomedicine*, Vol.7, 344-357 (2003).

[9] Shiraishi J., Katsuragawa *S.*, Ikezoe *A. et al.*, “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,” AJR,Am. J. Roentgen. 174(1), 71–74 (2000).

[10] Hardie R. C., S. K. Rogers, T. Wilson and A. Rogers,"Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs," Medical Image Analysis, vol. 12, pp. 240-258, 2008.

[11] Xu Y., Lee M. E., Boroczky L. *et al.*, "Comparison of Image Features Calculated in Different Dimensions for Computer-Aided

Diagnosis of Lung Nodules," Medical Imaging, Proc. of SPIE vol. 7260, 72600Z-1, 2009.

[12] Sun T., R. Zhang, J. Wang *et al.*, " Computer-Aided Diagnosis for Early-Stage Lung Cancer Based on Longitudinal and Balanced Data," PLOS ONE, Vol. 8 , Issue 5, pp.1-6, 2003.

[13] Eltoukhy M. M., I. Faye and B. B. Samir," Breast cancer diagnosis in digital mammogram using multiscale Curvelet transform," Computerized Medical Imaging and Graphics vol. 34 pp. 269–276, 2010.

[14] Schilham A. M. R., B. V. Ginneken and M. Loog,"A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database," Medical Image Analysis, vol. 10, pp. 247-258, 2006.

[15] Al Gindi A., Rashed E., Sami M., "Development and Evaluation of a Computer-Aided Diagnostic Algorithm for Lung Nodule Characterization and Classification in Chest Radiographs using Multiscale Wavelet Transform" *J Am Sci*; 9(x). (ISSN: 1545-1003) 2013.

[16] Yadav N. G., "Detection of Lung Nodule using Content-Based Medical Image Retrieval," International Conference on Advanced Engineering & Technology, Pune, ISBN: 978-81-925751-1-7, pp. 15-18, 2012.

[17] Zhu Y., Tan Y., Hua Y. *et al.*,"Feature Selection and Performance Evaluation of Support Vector Machine (SVM)-Based Classfier for Differentiating Benign and Malignant Pulmonary Nodules by Computed Tomography," Journal of Medical Imaging, vol. 23, no. 1, pp.51-65, 2010.

[18] Guo Q., M. Xu and J. Zhang,"A Novel Fast Marching Segmentation Algorithm for Pulmonary Nodules in Chest Radiographs," Yi Peng, Xiaohong Weng (Eds.), IFMBE Proceedings 9, Springer-Verlag Berlin Heidelberg, pp. 225–228, 2008.

[19] Chen S., K. Suzuki and H. MacMahon, "Development and evaluation of a computer-aided diagnostic scheme for lung nodule detection in chest radiographs by means of two-stage nodule enhancement with support vector classification," Med. Phys. Vol. 38, no.4, pp. 1844-1858, 2011.

[20] Rashed E. A., I. A. Ismail and S. I. Zaki," Multiresolution mammogram analysis in multilevel decomposition," Pattern Recognition Letters vol. 28, pp. 286-292, 2007.

4/15/2014