Lung cancer detection using X-ray and pattern recognition

Reported to Helwan university
faculty of computers and informations

BY

Mohamed Ahmed   Mohamed Said
Mohamed Bayoumi   Mostafa Bayoumi
Mohamed Samy   Omar Ahmed

Under the supervision of

A.Prof Waleed A.Yousef
Prof Amr Ghonaim

Abstract

Cancer is a leading cause of death worldwide. Lung cancer is a type of cancer that is considered as one of the most leading causes of death globally. This report aims to review existing approaches to the Lung cancer Detection and segmentation in CXRs images, highlighting the main differences between the used methods and also contains a comparative study will elaborate difference between 4 strategies based on 3 different classifiers (SVM & KNN & XGBOOST) and Wavelet transformation by it's all families as a Feature Extraction. and Wavelet transformation by it's all families as a Feature Extraction.

1 Introduction

Cancer is a disease that is referred to as the number one cause of death worldwide. According to the World Health Organization (WHO), cancer was the reason for 7.4 million deaths (13% of all deaths) occurred in 2008. More than 70% of all cancer deaths occur in middle income nations. It is projected that deaths caused by cancer will be growing to reach 13.1 million by the year 2030 . Lung cancer is a form of cancer that has become a significant cause of death globally . According to GLOBOCAN Project – which is carried out by the International Agency for Research in Cancer (An agency of the WHO) – lung cancer has accounted for 12.7 % of the total cancer incidents and 18.2 % of the total mortality caused by cancer, both rates are more than any other form of cancer.Figure 1 shows a summary for the cancer incident and mortality rates in both sexes worldwide as it was published by GLOBOCAN project

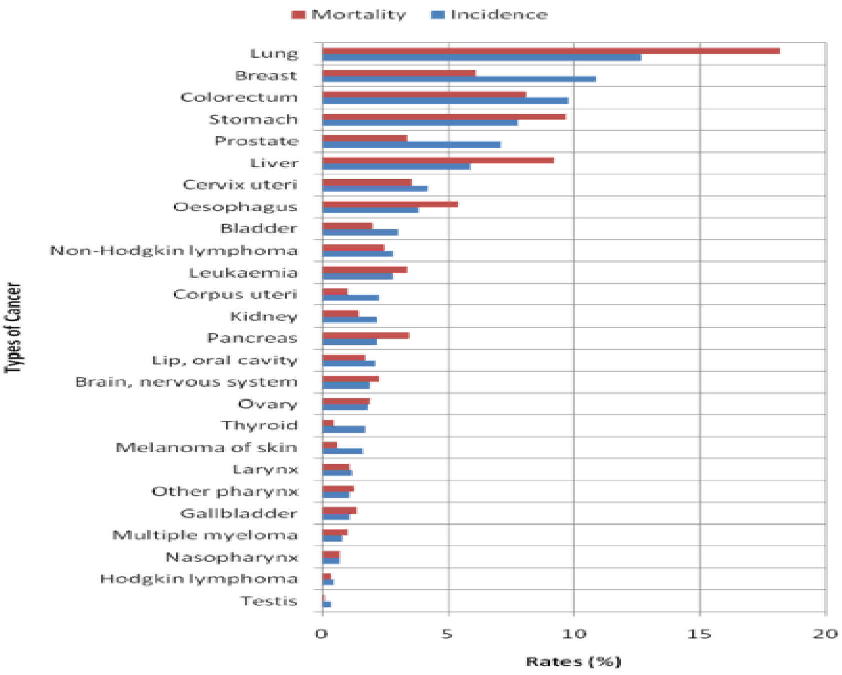


Figure 1: Cancer incidence and mortality rates

2 Nodule Detection Strategies

This section show different strategies and features used to detect and classify lung nodule.

2.1 Multi-scale Nodule Detection in Chest Radiographs[1]

In this work, a novel computer algorithm for nodule detection in chest radiographs is presented that takes into account the wide size range for lung nodules through the use of multi-scale image processing techniques. The method consists of:

- Lung field segmentation with an Active Shape Model.
- Nodule candidate detection by Lindeberg's multi-scale blob detector and quadratic classification.
- Blob segmentation by multi-scale edge focusing.
- k Nearest neighbor classification.

Experiments on the complete JSRT database show that by accepting on average 2 false positives per image, 50.6% of all nodules are detected. For 10 false positives, this increases to 69.5%.

Result, area under the ROC curve $A_z = 0.568$, result showed in term of FROC in Figure 2

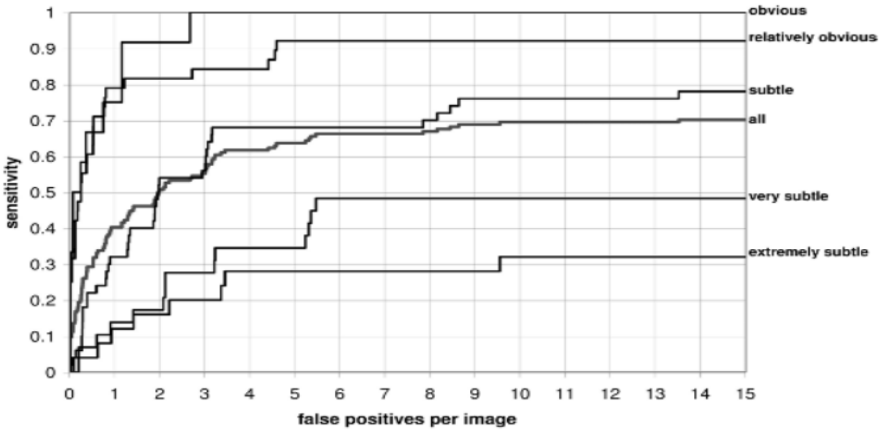


Figure 2: FROC curves of the system on the complete JSRT database, and per category

2.2 Optimal image feature set for detecting lung nodules[2]

This strategy consists of four processing steps:

1. location of tumor candidates by using adaptive ring filter.
2. extraction of the boundaries of tumor candidates by using IRIS filter,SNAKES algorithm is applied to the output image of the Iris filter to obtain the boundary of the tumor candidate.
3. extract feature by wavelet transformation , spatial difference and multi-scale image
4. discrimination between the normal and the abnormal regions is performed using a statistical method based on the Maharanobis distance measure.

Result,Figure 3 shows the relationship between the area under the ROC curve (A_z) and the number of features selected by the forward step-wise selection method

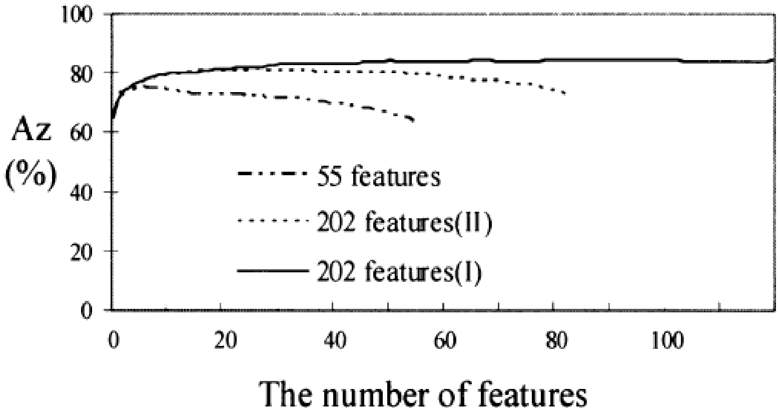


Figure 3: The relationship between the number of features and the area A_z

2.3 Detection of Lung Nodule Candidates in Chest Radiographs[3]

An innovative operator, called sliding band filter (SBF), is used for enhancing the lung field areas. In order to reduce the influence of the blood vessels near the mediastinum, this filtered image is multiplied by a mask that assigns to each lung field point an a priori probability of belonging to a nodule. The result is further processed with a watershed segmentation method that divides each lung field into a set of non-overlapping areas. Suspicious nodule locations are associated with the regions containing the highest regional maximum values.

Result,The algorithm proposed in this strategy was evaluated on a publicly available database, the JSRT database, result showed in term of FROC in Figure 4

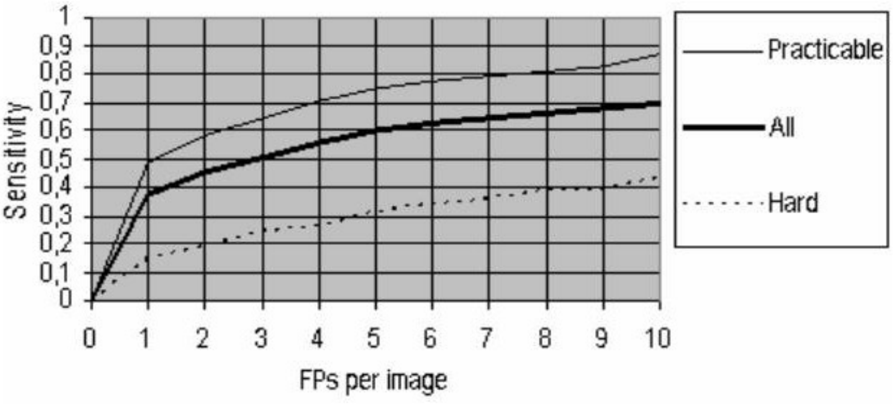


Figure 4: FROC curves for the complete JSRT database, showing the sensitivity of our method for all nodules, for the practicable nodules (subtlety levels 3, 4 and 5) and for the hard nodules (subtlety levels 1 and 2)

2.4 Lung Cancer Classification Using Image Processing[4]

This strategy consists of five steps:

2.4.1 Pre-Processing Of Images

- Generally during the scanning, the quality of image is affected by different artifacts due to non uniform intensity, variations, motions, shift, and noise. Thus, the pre-processing of image aims at selectively removing the redundancy present in scanned images without affecting the details, that play a key role in the diagnostic and analysis process

2.4.2 Lung Field Segmentation

- To restrict the segmentation area as well as to remove the other bony structures (like shoulder bones), segmentation of lung fields is essential. In lung fields segmentation, lung field masks have been prepared manually by deriving the peripheral pixel co-ordinates. Mask is a logical image denoting field area with logical 1 values and rest of the area with logical 0 values. Lung fields are separated from the rest of the image portion by multiplying mask with the median filtered image.

2.4.3 Lung nodule Segmentation

- Thresholding technique is applied on the separated lung fields images. The valley point value between the two peaks of the histogram is used as threshold value for segmentation of nodule.

- Geometrical features like area, diameter, perimeter, and irregularity index have been estimated from the separated lung nodules

- Contrast features are again classified under two categories, first order statistic and second order statistic.

- Texture related features like average gray level, standard deviation, smoothness, third moment uniformity and entropy are estimated using Gray Level Co-Occurrence Matrix technique (GLCM).

2.4.4 ANN based classification

- Further these estimated features are applied as input pattern to an expert system, which is designed to test the effectiveness of the input features so as to discriminate the lung cancer images. Artificial Neural Network (ANN) theory and practice suggest that, in a diagnostic application, the network should be trained with a balanced mixture of inputs from each diagnostic class.

Result,The diagnostic results obtained are found to be very promising. As high as 83% accuracy in classification is achieved using training data sets of reasonable size. Classification accuracy is improved as the numbers of training samples are increased.

2.5 Computer Aided Diagnosis System based on Machine Learning Techniques for Lung Cancer[5]

The proposed system was tested using a data set that consists of 167 chest radiography that contains 181 lung nodules the system utilized adaptive distance-based threshold algorithm for nodule segmentation, after that, features were computed for each nodule using geometric features, intensity features and gradient features. Lastly, a Fisher linear discriminant classifier was used to classify the computed features.
Result, The system could detect 78.1% of those nodules.

2.6 A Computer Aided Pulmonary Nodule Detection System Using Multiple Massive Training SVMs[6]

2.6.1 Lung Segmentation

To eliminate false positives that might occur outside of the lung, it is important to obtain a very accurate segmentation of the lung boundaries.
In this work, a double localizing region-based active model is employed for segmentation of lung filed in chest radiographs.

Firstly, we use a self-adaptive thresholding technique to categorize the image into three parts: visible lung, air and other human tissue.

Next, by fitting the average locations of the midpoint of left and right ribcage edge to a straight line using least square solution, the midline of lung area is determined.

Then double localizing regions are define and initialized based on the parameter of midline of lung area.

2.6.2 Potential nodule candidate detection

Given the fact that a nodule is generally either spherical or has local spherical elements, while a blood vessel is usually oblong. The eigenvalues of Hessian matrix of an image, which provides information about the shape of the considered region, can be used as a tool to identify any region which has a spherical structure (where a potential nodule may happen to occur) in a chest radiograph.

2.6.3 False positive reduction

A two stage classifier method is developed to address the problem of false positive reduction in lung nodule detection. Firstly, a rule-based classifier is employed to quickly remove obvious FP outliers so that their influence on the training of the second classifier was eliminated. Then a filter termed as multiple massive training supported vector machine MTSVM is developed to further separate nodules from nonnodule candidates.

2.6.4 Overall system performance

Shows the experimental results.Figure 5, it can be seen that the overall sensitive of the proposed CAD scheme can approach to 85% while that of an individual SVM based CAD scheme and a FLD classifier based CAD scheme can only approach to 83% and 81%, respectively. It can also be noticed that, with the proposed CAD system, at an overall sensitivity of 85%, the false positive was reduced from 12 to 4 positives per image. However, with an individual SVM based CAD scheme or a FLD classifier based CAD scheme, the false positive was only reduced from 12 to 6 per image at an overall sensitivity of 83% and of 81%, respectively. These results suggest that the proposed scheme was superior to others in FPs reduction in lung nodule detection on chest radiograph.

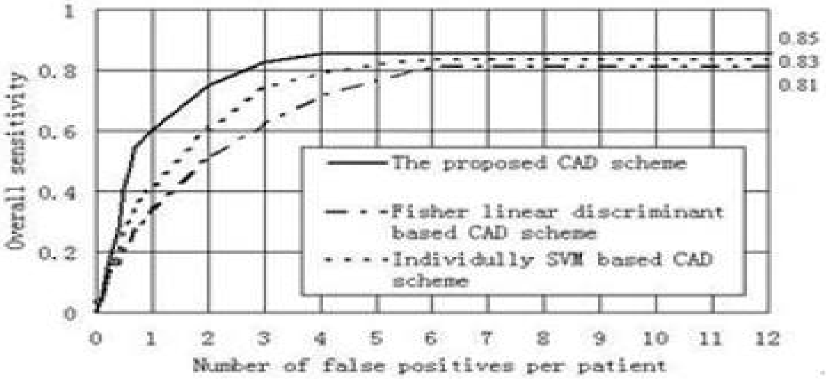


Figure 5: Evaluation of overall system performance by a comparison of the proposed CAD scheme with individual SVM classifier based CAD scheme and a Fisher linear discriminant (FLD) classifier based CAD scheme using the free response receiver operating characteristic (FROC) curves.

2.7 A Computer Aided Diagnosis System for Lung Cancer based on Statistical and Machine Learning Techniques[7]

2.7.1 Pre-processing

Two pre-processing techniques are employed in this CAD system.

- The first technique is histogram equalization to enhance the contrast of the image so the intensity levels (which will be in the range [0 1]) would generate a flat histogram.
- The second technique is image filtering. Laplacian filter is utilized to spotlight the rapid intensity change in regions within an image.

2.7.2 Feature Extraction

The transform of a signal is regarded as an alternative way of presenting the signal. And it also saves the important information contained in the signal. Wavelets employ various sets of fundamental functions to allow for the decomposition of continuous and discrete signals. Wavelet Transform provides a time-frequency description of the signal.

2.7.3 Feature Selection

It is essential to lower the number of the coefficients generated by the wavelet. This can be done by determining those coefficients which contains relevant information only.

The selected features are computed in two stages, which are calculating a statistical energy as well as a statistical metric. In both stages, only features less than a specific threshold are selected.

2.7.4 Classification

A classifier which is a combination of K-means modified algorithm and K-Near Neighbor (K-NN) is used.

2.7.5 Results

The results proves the capabilities of the CAD system in lung cancer classification (Normal Vs. Abnormal and Benign Vs. Malignant). A 99.15 % and 98.70% accuracy rates have been achieved for the normal vs. abnormal and benign vs. malignant experiments respectively with zero false positives and false negatives.

2.8 A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database[8]

2.8.1 Preprocessing

The efficacy of the detection and segmentation stages in the CAD scheme is boosted by preprocessing of the input images. For both steps, images are down-sampled to reduce the computational effort needed and filtered to obtain comparable contrast throughout each image. Also, the detector output is confined to a region of interest

2.8.2 Lung field segmentation

Local normalization Ofttimes on a chest X-ray, a nodule has a poor contrast to the background. Not only can a nodule be intrinsically hard to distinguish, being very small, or having a verylow density, but frequently the nodule is partly obscured by structures, such as ribs and vessels. By local normalization(LN) filtering, a global equalization of contrast throughout an image is achieved. This filtering also normalizes edge strengths, which enhance the performance of the blob detector and improves the process of segmentation.

2.8.3 Features Extraction

this strategy choose features from a multi-scale Gaussian filterbank and a small number of specific features that are readily calculated from the blob detector scheme.The Gaussian filterbank consists of all Gaussian derivatives from 0th to 2nd order for 4 different scales (r = 1,2,4,8 pixels). A multi-scale filterbank is chosen to account for the spread in blob sizes

2.8.4 Classification

The classifier used is a k nearest neighbors (kNN) classifier which searches the feature space to find the k nearest neighbors of an object among all nodule candidates from all cases in the database.

2.8.5 Result

in Figure 6, Figure 7 the final performances of the four CAD schemes are shown as FROC curves for the JSRT database, measuring sensitivity (overall and for the hard and practicable cases separately) as a function of the average number of false positives per image. Indeed the graphs show that it is much harder to detect nodules of the hard category than those of the practicable class, which reflects the experience of the radiologists.

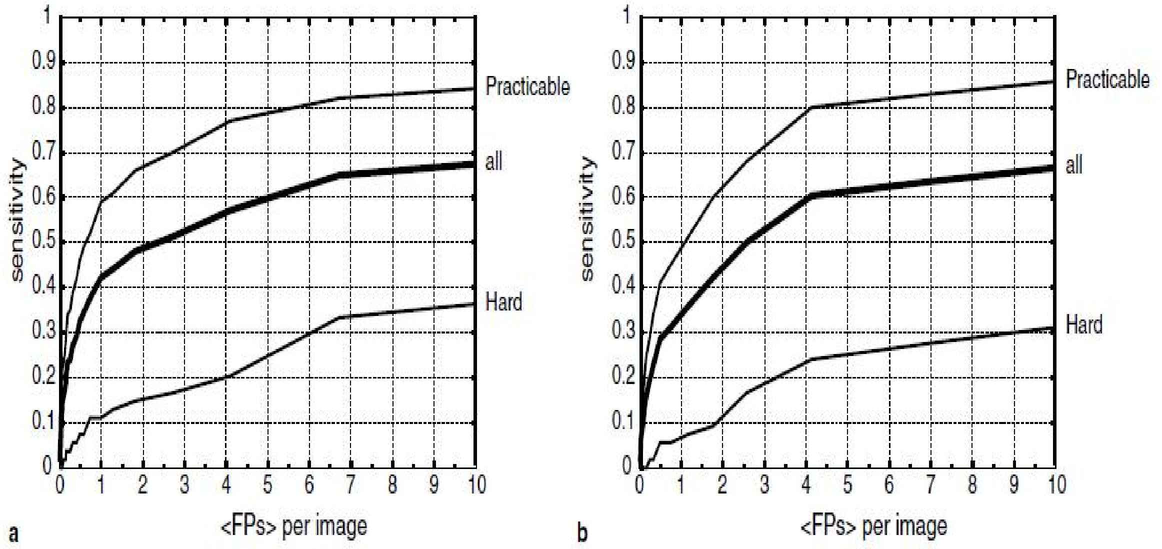


Figure 6: FROC curves of the system for the complete JSRT database, showing the sensitivity for all nodules, for the practicable nodules, and for the hard nodules; (a) for the basic scheme and (b) for the scheme with segmentation

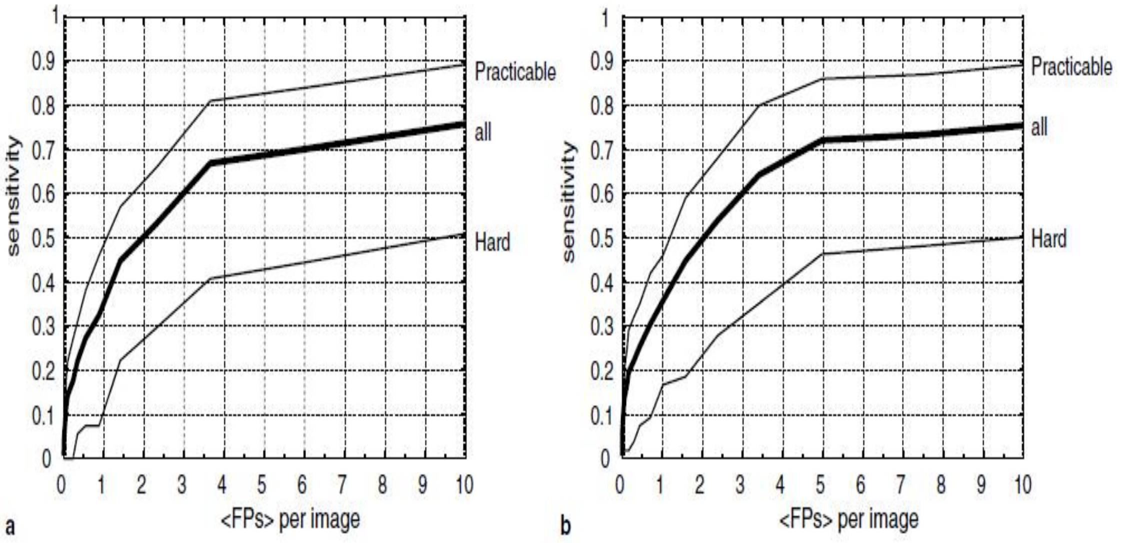
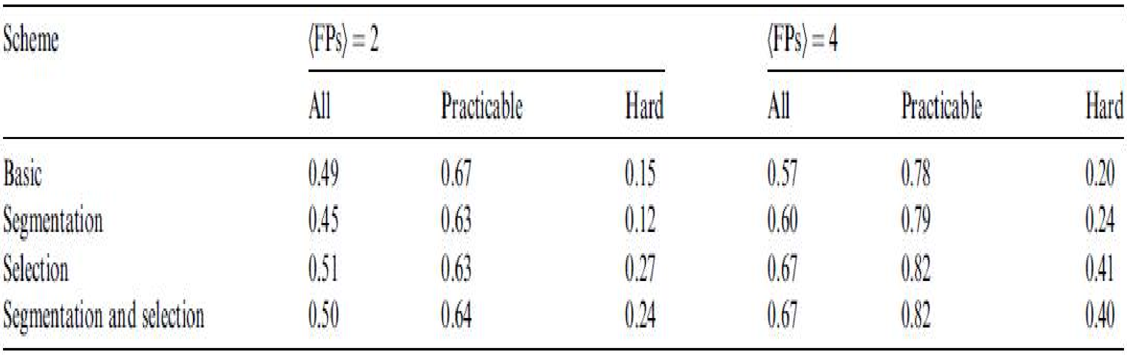


Figure 7: FROC curves of the system for the complete JSRT database, showing the sensitivity for all nodules, for the practicable nodules, and for the hard nodules; (a) for the scheme with selection and (b) for the scheme with selection and segmentation

The sensitivities of the CAD schemes at on average two and four false positives (FPs) per image are listed in Figure 8. These two values seem reasonable for a CAD system for nodule detection in radiographs; a system generating too many false positives would probably not be used by a radiologist, because it then significantly increases his workload, while the probability that the system is wrong on a per marker basis is very large. The numbers in Figure 8 show that the addition of the first selection step is always advantageous. This enhancement is most apparent for the hard cases. The addition of the segmentation stage does not have a clear positive effect on the sensitivity of the scheme at these two operating points. These results prompt us to choose the selection scheme as our CAD system of preference.



Scheme	⟨FPs⟩ = 2			⟨FPs⟩ = 4		
	All	Practicable	Hard	All	Practicable	Hard
Basic	0.49	0.67	0.15	0.57	0.78	0.20
Segmentation	0.45	0.63	0.12	0.60	0.79	0.24
Selection	0.51	0.63	0.27	0.67	0.82	0.41
Segmentation and selection	0.50	0.64	0.24	0.67	0.82	0.40

Figure 8: Performances of the four CAD schemes expressed as sensitivity when accepting on average two and four false positives per image

2.9 Nodule Detection in Chest Radiographs [9]

This strategy employ an intensity based blob detection algorithm, Lindeberg's multi-scale blob detection scheme.The idea of this Laplacian of the Gaussian (LoG) based algorithm is to extract blobs by detecting scale-space maxima.
Using 154 images, this algorithm missed 10 positives before feature extraction process with an average of 164 false positives per image.
As a result of the high number of false positives,A different methodology to extract possible nodule regions is used.
Rather than using intensity based methods, we study (CI) based methods which are proven to be the best nodule candidate detectors throughout different types of detectors.CI filters use the directions of the gradient vectors in circular regions which have brightness values that decreases from center to edges.In these regions the gradients point towards the center of the object which results with a high CI value. 5 variations of CI filters: Convergence Index, Adaptive Ring Filter, Sliding Band Filter, IRIS Filter and Weighted Convergence Index (WCI) filters is studied.WCI is different than regular CI filter in two ways.

First, it operates on three different levels instead of one radius and then the weights are different for these levels, the highest being the closest to the center.

After detection and binarization of WCI image, we remove very small and large components which cannot be nodule by morphologic opening operator.Small components are described as the regions which have less connected pixels than 20 and large components are the ones which have more connected pixels than 400.

Then we found regions that are not circular using two operators, ratioAxis and areaRatio. While the former found the elongated regions, the latter looks the ratio between area and axis lengths.
To identify nodule boundaries, an adaptive distance based threshold algorithm is applied to each nodule candidate.To eliminate overlapping nodule candidates, the candidate with the highest WCI value is kept while others are removed.

there is a wide range of feature set to describe each nodule candidate For each nodule candidate,A total of six sets of features is extracted ,Position features,Texture features useful for classification. Intensity features: Several values are calculated to describe the gray level distribution of the pixels inside and outside of the nodule region. Gaussian features: Since we do not have a prior knowledge about the size of the nodules, we use the multi-scale Gaussian filter, Gradient features: Except the value that comes from the Entropy of the Gradients, all of the computed values, Gradient Magnitude, Radial Deviation and Radial Gradient. To our knowledge, Entropy of the Gradients has not been used in the literature before. Detector features: These features are calculated using Hessian matrix of LCE image.

Result,CI filter results with 42 False Positives (FP) per image but missed 30 nodules on resampled images. Only 1 nodule out of all missed is related with threshold, while others are not visible even in CI results. In the second study with enhanced images they missed 8 nodules but number of FPs increased to 98. IRIS filter enhanced many structures but not nodules, resulting with too many fps per image and doesn't seem promising. ARF on resampled images results with 13 missed nodules and an average of 95 false positive per image. Using enhanced image instead of resampled image does not provide any improvement on results. they missed 14 nodules and get 117 fps per image on our study with ARF on enhanced image. Using global thresholding technique on SBF as they did on other filters, they missed 23 nodules with an average of 88 fps. Local thresholding, in SBF reduces the missed nodules from 23 to 16 but increases fp per image from 88 to 115. they perform their final experiments on WCI filter. Using enhanced images as input to WCI filter, they missed 6 nodules and got 98 fps per image. Since they got optimal sensitivity missed nodule rate using WCI filter.

2.10 Classification of Chest Lesions with Using Fuzzy C-Means Algorithm and Support Vector Machines[10]

2.10.1 Fuzzy C-Means Algorithm

the fuzzy diagnosis concept is widely applied . Fuzzy classifiers have been proposed to deal with classification tasks in presence of uncertainty. Fuzzy c-means (FCM) is one of these methods. It is an algorithm of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition.

2.10.2 SVM Classification

Support vector machines (SVM) represent a classifier that has been successfully used for chest lesions classification. Moreover, we possess labeled data and it is expected that a supervised classifier achieves a good accuracy.

2.10.3 Preprocessing

Preprocessing may lead to better results because it allows more flexibility. For example, the contrast feature is almost meaningless in discriminating normal and lesion pixels, since lesions can occur in regions of both high and low contrast. However, combining this feature with image enhanced intensity can be used for much more effective discrimination

2.10.4 Feature Extraction and Selection

The purpose of features extraction and selection is to reduce the original data set by measuring certain properties that distinguish one input pattern from another pattern. The extracted feature should provide the characteristics of the input type to the classifier by considering the description of relevant properties of the image into a feature space.
Feature Selected are (size, circularity, x-fraction, y-fraction, skewness, kurtosis, homogeneity, correlation)

2.11 Geometrical and texture features estimation of lung cancer and TB images using chest X-ray database [11]

Most of the lung cancers start in the lining of the bronchi. Less often, cancers begin in the trachea, bronchioles, or alveoli. Lung cancers are thought to develop over a period of many years. As a cancer develops, the cancer cells may produce chemicals that cause new blood vessels to form nearby. These new blood vessels nourish the cancer cells, which can continue to grow and form a tumour large enough to see on X-rays. Cells from the cancer can break away from the original tumour and spread to other parts of the body. As noted earlier, this process is called metastasis. Mainly, the lung cancers are classified as Small-Cell type of Lung Cancer (SCLC), and Non-Small-Cell type of lung cancer (NSCLC) (Figure 9). Usually, SCLC arises at alveolar level or at terminal bronchial level, and seen to be more scattered in nature on X-ray. NSCLC arises in the larger, more central bronchi; tends to spread locally; metastasises somewhat larger than the other patterns, but its rate of growth in its site of origin is usually more rapid than that of other types.

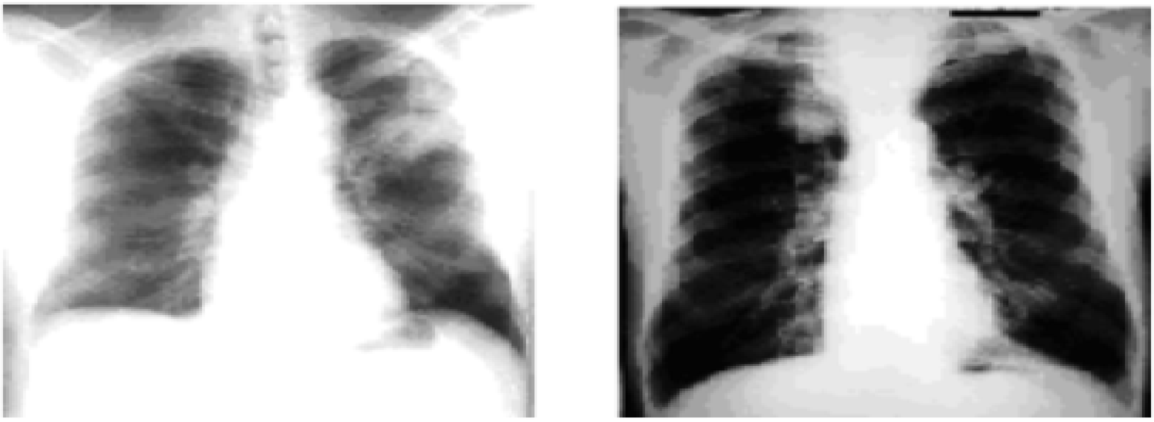


Figure 9: SCLC and NSCLC images

Tuberculosis (TB), though no longer the ubiquitous and sinister threat of past decades, is still a problem that must be borne in mind, particularly with immigrants from developing countries and in the immune-suppressed patients. The primary form of TB infection, which is used to be seen almost exclusively in children, is now also being seen in older patients. The primary TB infection appears on X-ray as a diffuse opacity representing a patch of consolidation in the lung field with increased striations extending towards the hilum where the enlarged glands show as rounded opacities. Pleural effusion is also a common manifestation of primary TB infection. The secondary or adult type of TB infection mostly affects the posterior segment of the upper lobe. On the Posterior–Anterior (PA) film, it appears as an area of shadowing near the lung apex often mottled in character. Figure 10 gives an idea of the spatial distribution of the abnormal areas within the chest radiographs in the TB database. Feature selection is a very important step in organising a classifier. Theoretical approach cannot be applied to determine the optimal combination of features, and the only way to select the optimal feature subset is to evaluate all possible combinations of the features. Moreover, sufficient numbers of test materials are necessary to evaluate the performance of each feature combination. It means that the number of combinations and the total amount of computation time become impractically huge. Therefore, Jun Wei, Yoshihiro Hagihara, Akinobu Shimizu, and Hidefumi Kobatake accepted heuristic algorithms such as a genetic algorithm, a forward stepwise and a backward stepwise selection technique to decide the optimal feature set.

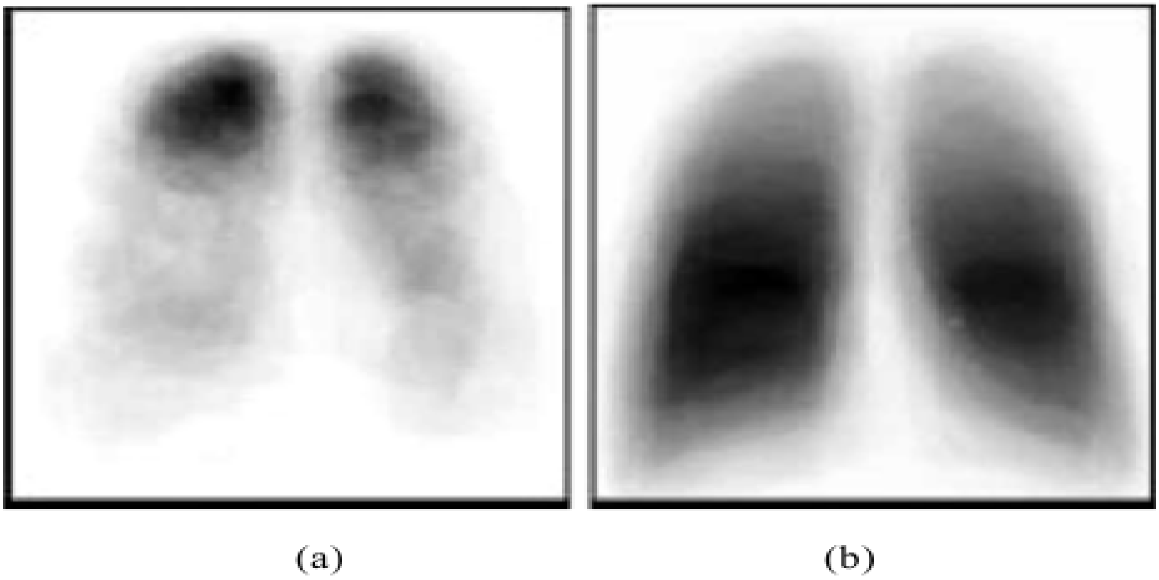


Figure 10: (a) Distribution of abnormal areas for the TB database and (b) distribution of abnormal areas for the Interstitial Disease (ID) database

the features are classified as

1. Geometric features

   Spreadness, Circularity, Area, Equivalent radius, Distance from the candidate point to the pulmonary hilum and Flatness are some of the geometric features. Such geometric features are calculated from the binary Suspicious Region (SR) using thresholding technique.

2. Texture or contrast features

   Generally, tumour region is brighter than its background on X-ray image. So, the contrast information can be used as features. Contrast features are again classified under two categories, first-order statistic and second-order statistic. In this work, such 10 kinds of features are calculated from SRs.

   2.1. First-order statistics features

   First-order statistics are calculated from histograms of the grey-scale values. The histograms are obtained from filtered images. Features calculated from each histogram include average grey level, standard deviation, contrast, skewness, kurtosis and entropy.

   2.2. Second-order statistics features

   Co-occurrence matrix method has been adopted to extract features of second-order statistics. They are obtained by using Haralick transformation. Co-occurrence matrices are obtained from the inner and the outer regions of each SR. Correlation, energy, homogeneity and contrast are the features computed by using co-occurrence matrix.

2.12 Computerized Detection of Lung Nodules by Means of “Virtual Dual-Energy” Radiography[12]

2.12.1 Abstract :

this system is using VDE(virtual dual energy) radiography for suppressing rips and clavicles to reduce false positives.

2.12.2 Methods:

CADe scheme for detection of lung nodules in CXRs consisted of four major steps:
1) segmentation of lung fields based on our multisegment active shape model (M-ASM);
2) two-stage nodule enhancement and nodule candidate detection;
3) segmentation of nodule candidates by use of our clustering watershed algorithm;
4) feature analysis and classification of the nodule candidates into nodules or non-nodules by use of a nonlinear support vector machine (SVM) classifier

2.12.3 Result

the VDE-based CADe scheme achieved a sensitivity of 85.0% (119/140) and the original CADe scheme achieved a sensitivity of 78.5% (110/140) at an FP rate of 5.0 FPs per image for the JSRT database . The performance of the CADe scheme (85% sensitivity with 5 FPs/image) provided a substantial improvement against the original CADe scheme (78% sensitivity with 5 Fps/image).

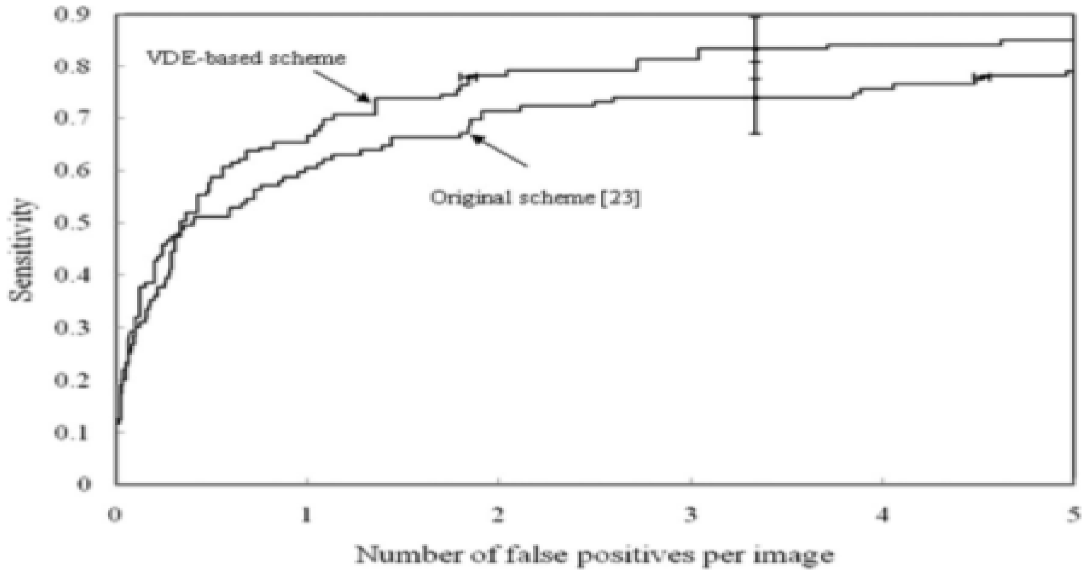


Figure 11: FROC curves indicating the improvement in the performance of our CADe schemes with SVM classifier by use of our VDE technology for the JSRT database. Error bars indicate 95% confidence intervals.

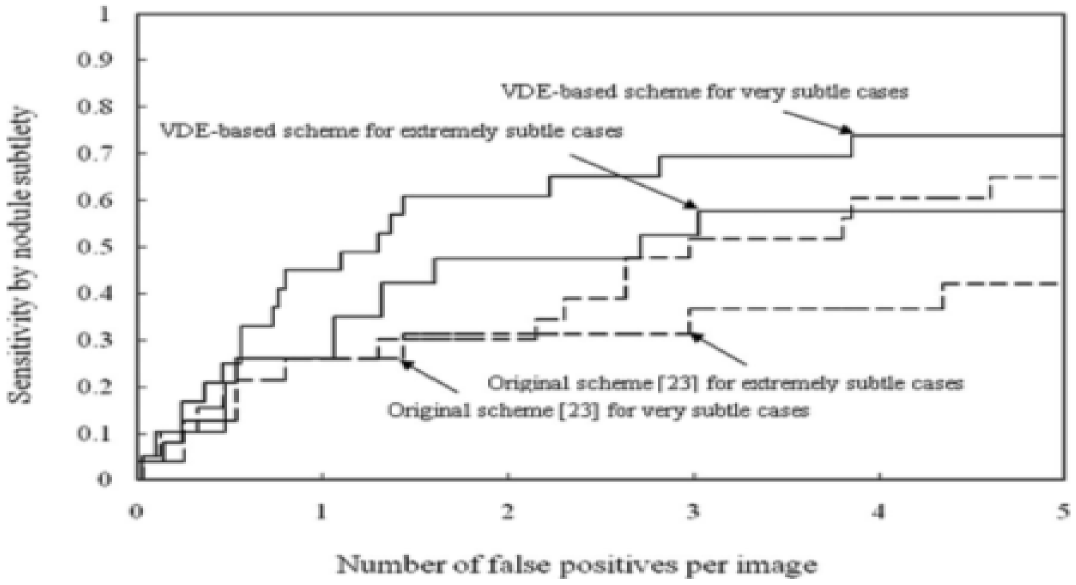


Figure 12: FROC curves indicating the performance of the VDE-based CADe scheme by nodule size for the JSRT database

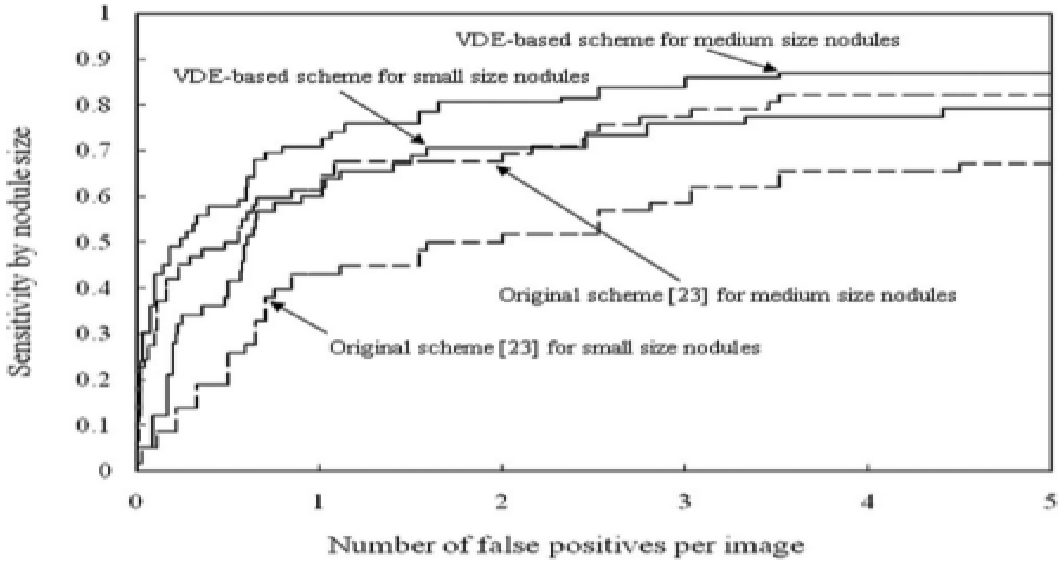


Figure 13: FROC curves indicating the performance of the VDE-based CADe scheme by nodule subtlety for the JSRT database.

2.13 On the Combination of Wavelet and Curvelet for Feature Extraction to Classify Lung Cancer on Chest Radiographs[13]

2.13.1 Feature extraction

Feature extraction is a significant step in CAD system. In this paper we utilize wavelet and curvelet separately and combined in order to investigate their performance and their contribution to the diagnosis of lung cancer. Both wavelet and curvelet are well known techniques

2.13.2 Feature selection

Feature selection is another important step m CAD system. The aim of this step is to remove the extracted features that are redundant and have no contribution to the classification process. After applying wavelet, curvelet or the fusion of both of them, the generated coefficients are subjected to a two-step feature selection, namely, statistical energy and statistical metric.

2.13.3 Classification

Cluster-k-Nearest Neighbor (C-k-NN) classifier is utilized in this CAD system. The C-k-NN is a classifier that merges two algorithms that are K-means modified algorithm and k-Nearest neighbor algorithm. The K-means algorithm is utilized to cluster the data into classes and sub-classes with a central point to represent each class and sub-class while k-Nearest Neighbor is used to classify new data by calculating the Euclidean distance between the center point of each class and the new data. With this algorithm, the advantage of k-means of calculating in less time and the advantage of k-NN, which is the accurate classification, are combined.

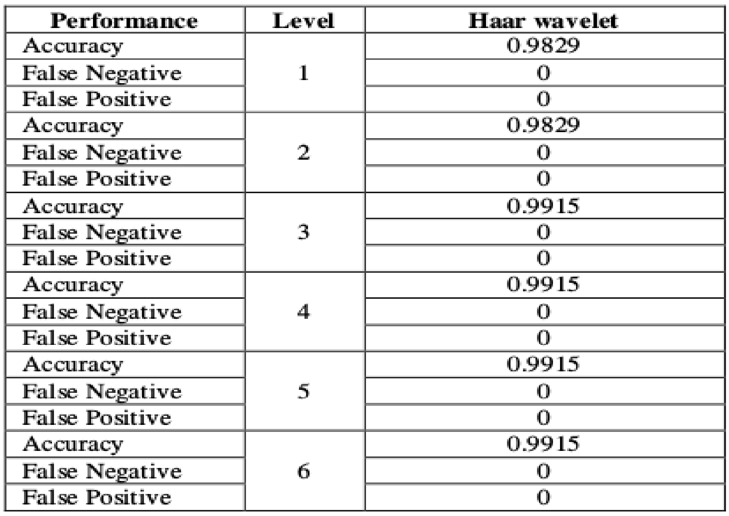
2.13.4 Result

In this paper, the CAD system is applied in two main experiments, classification of normal from abnormal cases and benign from malignant cases. In each experiment, different scenarios of combining the coefficients of wavelets and curvelet are applied as follows:

- Wavelets transform (haar function).
- Combination of the coefficients of two levels of haar wavelet.
- Combination of haar and db1 coefficients.
- Curvelet transform at scales 2-7.
- Combination of haar wavelet and Curvelet coefficients. For the combination of different wavelet levels or wavelet and curvelet, the coefficients were extracted separately and then combined in one matrix and randomized. After that, they were subjected to the feature selection and classification methods.

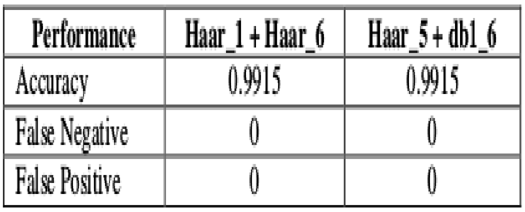
In the experiments, the sub-images (128x128) are used in the evaluation of the CAD system. Those images are divided into training and testing sets. The training set was used to train the classifier while the testing set was used to assess the classification performance.

A. Normal Vs. Abnormal For this experiment, Tables (II - V) show the obtained results.



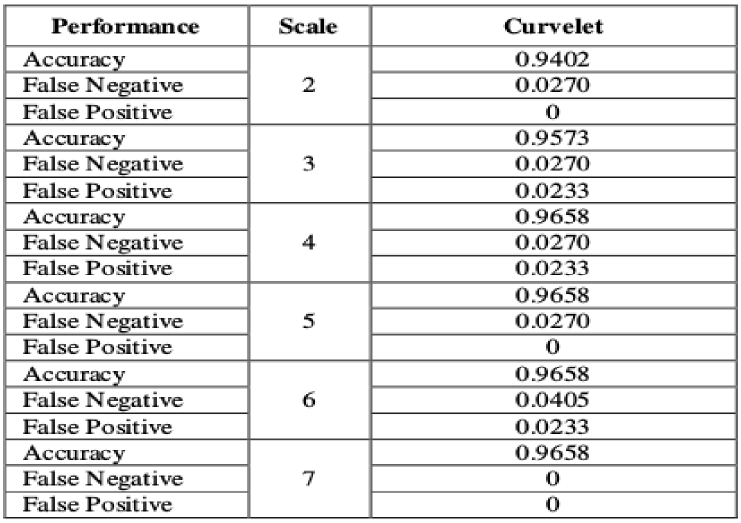
Performance	Level	Haar wavelet
Accuracy	1	0.9829
False Negative		0
False Positive		0
Accuracy	2	0.9829
False Negative		0
False Positive		0
Accuracy	3	0.9915
False Negative		0
False Positive		0
Accuracy	4	0.9915
False Negative		0
False Positive		0
Accuracy	5	0.9915
False Negative		0
False Positive		0
Accuracy	6	0.9915
False Negative		0
False Positive		0

Figure 14: NORMAL VS . ABNORMAL CLASSIFICATION WITH WAVELET



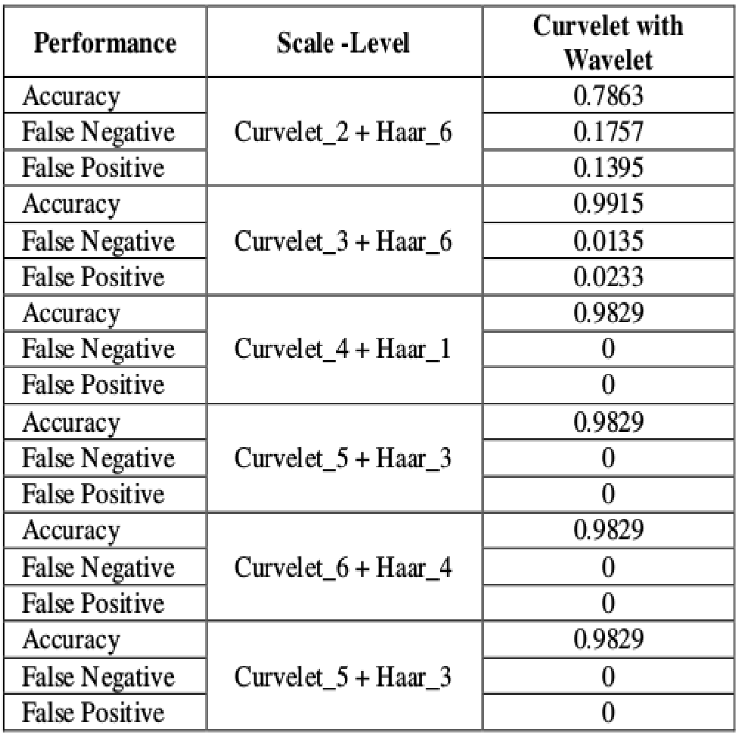
Performance	Haar_1 + Haar_6	Haar_5 + db1_6
Accuracy	0.9915	0.9915
False Negative	0	0
False Positive	0	0

Figure 15: NORMAL VS . ABNORMAL CLASSIFICATION WITH COMBINED WAVELETS



Performance	Scale	Curvelet
Accuracy	2	0.9402
False Negative		0.0270
False Positive		0
Accuracy	3	0.9573
False Negative		0.0270
False Positive		0.0233
Accuracy	4	0.9658
False Negative		0.0270
False Positive		0.0233
Accuracy	5	0.9658
False Negative		0.0270
False Positive		0
Accuracy	6	0.9658
False Negative		0.0405
False Positive		0.0233
Accuracy	7	0.9658
False Negative		0
False Positive		0

Figure 16: NORMAL VS . ABNORMAL CLASSIFICATION WITH CURVELET

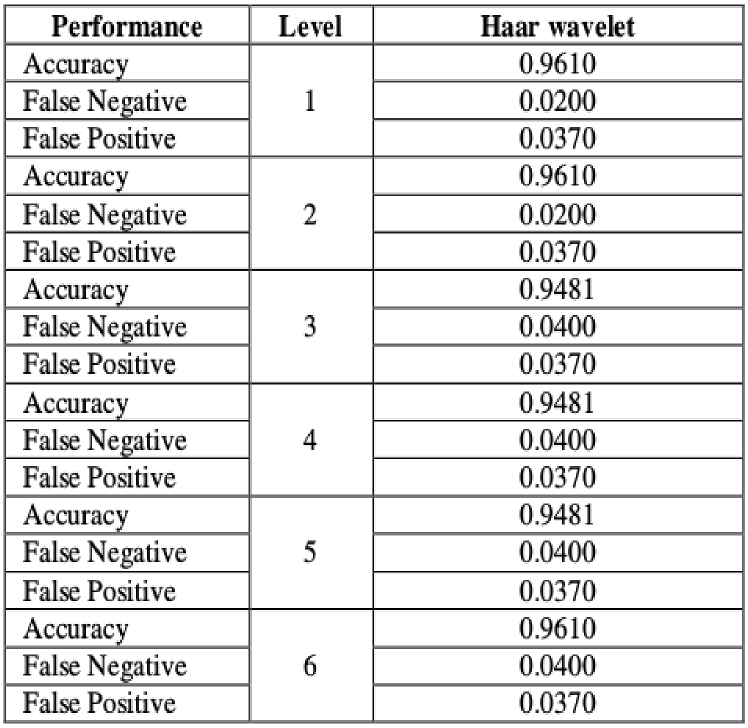


Performance	Scale -Level	Curvelet with Wavelet
Accuracy	Curvelet_2 + Haar_6	0.7863
False Negative		0.1757
False Positive		0.1395
Accuracy	Curvelet_3 + Haar_6	0.9915
False Negative		0.0135
False Positive		0.0233
Accuracy	Curvelet_4 + Haar_1	0.9829
False Negative		0
False Positive		0
Accuracy	Curvelet_5 + Haar_3	0.9829
False Negative		0
False Positive		0
Accuracy	Curvelet_6 + Haar_4	0.9829
False Negative		0
False Positive		0
Accuracy	Curvelet_5 + Haar_3	0.9829
False Negative		0
False Positive		0

Figure 17: NORMAL VS . ABNORMAL CLASSIFICATION WITH COMBINED WAVELET AND CURVELET

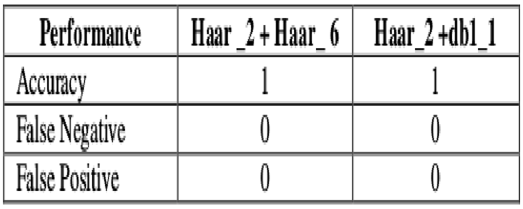
In this experiment, haar wavelet reached a maximum accuracy of 99.15% while curvelet reached 96.58%. However, after combining both wavelet with curvelet, the highest accuracy achieved is 99.15 which is better than the accuracy achieved by curvelet alone. Similar increase in the accuracy is also shown in the other experiment.

Benign vs. malignant For this experiment, Tables (VI-IX) show the obtained results.



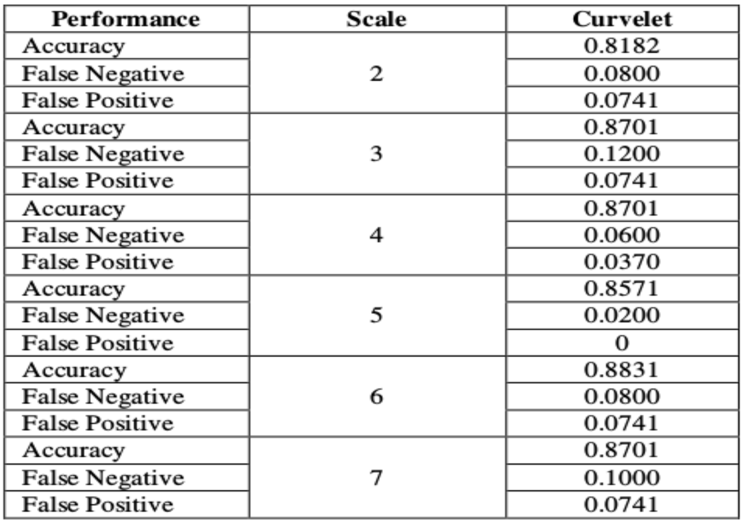
Performance	Level	Haar wavelet
Accuracy	1	0.9610
False Negative		0.0200
False Positive		0.0370
Accuracy	2	0.9610
False Negative		0.0200
False Positive		0.0370
Accuracy	3	0.9481
False Negative		0.0400
False Positive		0.0370
Accuracy	4	0.9481
False Negative		0.0400
False Positive		0.0370
Accuracy	5	0.9481
False Negative		0.0400
False Positive		0.0370
Accuracy	6	0.9610
False Negative		0.0400
False Positive		0.0370

Figure 18: BENIGN VS . MALIGNANT CLASSIFICATION WITH WAVELET



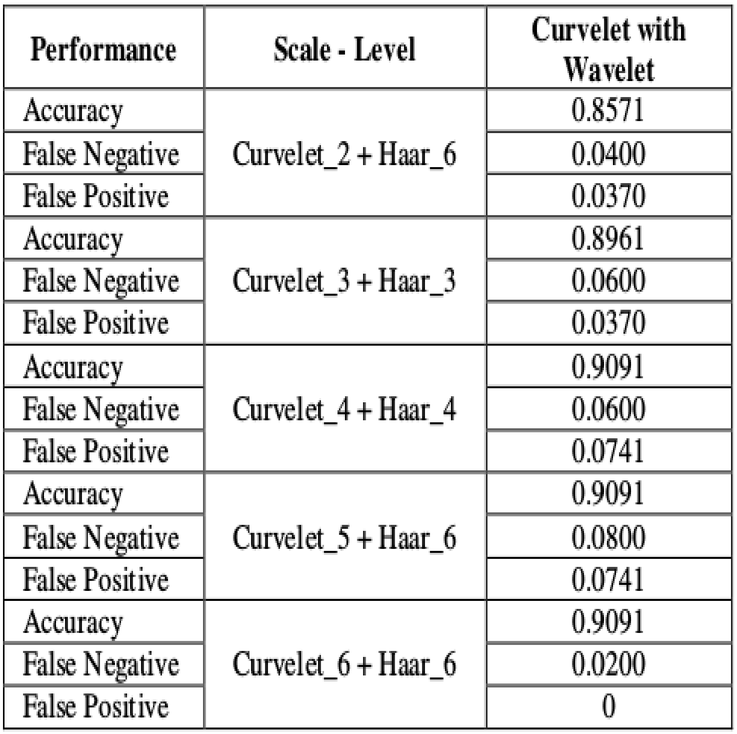
Performance	Haar _2 + Haar_ 6	Haar_2 +db1_1
Accuracy	1	1
False Negative	0	0
False Positive	0	0

Figure 19: BENIGN VS . MALIGNANT CLASSIFICATION WITH COMBINED WAVELETS



Performance	Scale	Curvelet
Accuracy	2	0.8182
False Negative		0.0800
False Positive		0.0741
Accuracy	3	0.8701
False Negative		0.1200
False Positive		0.0741
Accuracy	4	0.8701
False Negative		0.0600
False Positive		0.0370
Accuracy	5	0.8571
False Negative		0.0200
False Positive		0
Accuracy	6	0.8831
False Negative		0.0800
False Positive		0.0741
Accuracy	7	0.8701
False Negative		0.1000
False Positive		0.0741

Figure 20: BENIGN VS . MALIGNANT CLASSIFICATION WITH CURVELET



Performance	Scale - Level	Curvelet with Wavelet
Accuracy	Curvelet_2 + Haar_6	0.8571
False Negative		0.0400
False Positive		0.0370
Accuracy	Curvelet_3 + Haar_3	0.8961
False Negative		0.0600
False Positive		0.0370
Accuracy	Curvelet_4 + Haar_4	0.9091
False Negative		0.0600
False Positive		0.0741
Accuracy	Curvelet_5 + Haar_6	0.9091
False Negative		0.0800
False Positive		0.0741
Accuracy	Curvelet_6 + Haar_6	0.9091
False Negative		0.0200
False Positive		0

Figure 21: BENIGN VS . MALIGNANT CLASSIFICATION WITH COMBINED WAVELET AND CURVELET

2.14 LUNG CANCER DETECTION USING SUPERVISED CLASSIFICATION WITH CLUSTER VARIABILITY ON RADIOGRAPHS DATA[14]

In this paper, a comparative classification approach has been integrated with modified image based features filtering and selection to increase system detection rate by minimizing false alarms in terms of false positive and false negative rates.

2.14.1 Image filtering and equalization

The dataset images are in raw form which need sufficient pre-processing phase. Firstly histogram equalization is used which balances the intensity level and adjusts the contrast of the images to achieve a flat histogram.

After a repeated process of intensity equalizing, the images are filtered using the Laplacian method to further omit the odd regions of the images.

2.14.2 Feature transformation and selection

For signal decomposition, wavelet transformation is used which is considered better than curvelet transformation based on scaling image dimension. Such transformation helps in storing significant details which are useful for classifying the images by utilizing the signal's time-frequency content .

The next phase is to reduce the dimensionality of the newly generated coefficients from the wavelet transformation. Here two algorithms Supervised Principal Component Analysis (SPCA) and statistical energy based metric selection are applied. Both are modified enough to eliminate the large number of features produced from the image and figure out few high variant attributes which can be classified to determine the normal as well the cancerous images .

2.14.3 Supervised clustering

Once the features are selected, prior to classification, supervised clustering is applied to the processed features to highlight the possible dissimilar groups available in the dataset. In this approach, varying sizes are used, ranging from two which is default for carrying binary classification up to 25-30 depending upon the dataset. Clustered based classification helps in increasing the detection rate of the patterns because it is applied on the clusters separately, which have maximum similarity index and enhances the precision of the system.

2.14.4 Ensemble classification with optimized parametric tuning

In this ensemble approach binary classification is used in combination with clustering, which enhances the input data representation in the form of groups of patterns having similar behavior irrespective of their class labels for Support Vector Machine (SVM) classification where as k-NN takes clusters which have patterns of same class labels. The classifier architecture is based on the cross validation method which means there will be an iterative procedure of training and testing to ensure that the system is quantified with misclassifications which eventually decreases the false alarms.

2.14.5 Result

This work focused on the ensemble approach of clustering and supervised classifier for detection of cancer alignments where as the SPCA and statistical metrics performed an optimized feature selection and transformation into few variant attributes. To enhance the acquired accuracy, parameter tuning is also involved to get the best from the combination of applied techniques and algorithms. SVM performed the classification with less features and minimum time cost to achieve 94% accuracy, whereas k- NN utilized more features, considerable time cost and reach approx 97.2% accuracy rate. Overall, the ensemble system based on SVM classifier managed to overcome earlier drawbacks in terms of time and processing consumption. Related parameters such as cluster size also affect SVM processing and performance and they need further optimization to be more effective for determining possible cancer cases.

3 Conclusion

As it known that preprocessing may lead to better results because it allows more flexibility , Active shape model technique is widely used in preprocessing , and for feature extraction Wavelet transformation has shown great results in different CAD systems. Both Support Vector Machine and K-Near Neighbor is most common used in classification and has shown good results with false positive located between 1 and 9 per image.

4 References

1. Multi-scale Nodule Detection in Chest Radiographs

   Arnold M.R. Schilham, Bram van Ginneken, and Marco Loog Image Sciences Institute, University Medical Center Utrecht, The Netherlands

2. Optimal image feature set for detecting lung nodules on chest X-ray images

   Jun Wei, Yoshihiro Hagihara, Akinobu Shimizu, Hidefumi Kobatake

3. Detection of Lung Nodule Candidates in Chest Radiographs

   Carlos S. Pereira, Hugo Fernandes1, Ana Maria Mendonwca, and Aurmelio Campilho

4. Lung Cancer Classification Using Image Processing

   International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012

5. Computer Aided Diagnosis System based on Machine Learning Techniques for Lung Cancer

   Hamada R. H. AI-Absi, Brahim Belhaouari Samir, Khaled Bashir Shaban, and Suziah Sulaiman

6. A Computer Aided Pulmonary Nodule Detection System Using Multiple Massive Training SVMs

   Zhenghao Shi1, Minghua Zhao1, Lifeng He4, Yinghui Wang1, Ming Zhang2 and Kenji Suzuki3

7. A Computer Aided Diagnosis System for Lung Cancer based on Statistical and Machine Learning Techniques

   Hamada R. H. Al-Absi1, Brahim Belhaouari Samir2*, Suziah Sulaiman1

8. A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database

   Arnold M.R. Schilham *, Bram van Ginneken, Marco Loog

9. Classification of Chest Lesions with Using Fuzzy C-Means Algorithm and Support Vector Machines

   Donia Ben Hassen1, Hassen Taleb1, Ismahen Ben Yaacoub2, and Najla Mnif2

10. Geometrical and texture features estimation of lung cancer and TB images using chest X-ray database

    S.A. Patil

11. Computerized Detection of Lung Nodules by Means of virtual Dual energy Radiography

    Sheng Chen* and Kenji Suzuki, Member, IEEE

12. On the Combination of Wavelet and Curvelet for Feature Extraction to Classify Lung Cancer on Chest Radiographs

    Hamada R. H. Al-Absi, Brahim Belhaouari Samir, Taha Alhersh and Suziah Sulaiman

13. Computer-aided Diagnosis for the Detection and Classifcation of Lung Cancers on Chest Radiographs:

    Junji Shiraishi, PhD, Hiroyuki Abe, MD, PhD, Feng Li, MD, PhD, Roger Engelmann, MS Heber MacMahon, MB, Kunio Doi, PhD

Contents