RESEARCH ARTICLE

Detection of Juxtapleural Nodules in Lung Cancer Cases Using an Optimal Critical Point Selection Algorithm

S Saraswathi[1]*, L Mary Immaculate Sheela[2]*

Abstract

Detection of lung cancer through image processing is an important tool for diagnosis. In recent years, image processing techniques have become more widely used. Lung segmentation is an essential pre-processing step for most (CAD) schemes. An automated system is proposed in this paper for identifying lung cancer from the analysis of computed tomography images by performing nodule segmentation using an optimal critical point selection algorithm (OCPS) which improves the detection of shape- and size-based juxtapleural nodules located at the lung boundary. A suspect area of nodule is obtained with the help of a bidirectional chain code (BDC) approach and the OCPS This algorithm is used to reduce the time consumption to detect the lung nodule and thereby reduce the computational complexity. Shape and size features are extracted for the area between two critical points to facilitate classification as nodule or non-nodule with the help of a support vector machine and random forest classifiers. This automated method was tested on computed tomography (CT) studies from the lung imaging database consortium (LIDC). The results are compared with the existing techniques using various performance measures such as precision rate, recall rate, accuracy and F-measure. The obtained experimental results indicate that the OCPS combined with a random forest classifier performs better in terms of performance evaluation metrics than existing approaches, with less requirement for computation.

Keywords: Bi directional chain Code- SVM classifier- RF classifier- optimal critical point

Asian Pac J Cancer Prev, **18 (11)**, 3143-3148

Introduction

Lung cancer represents a major health problem. Cancer cells can be carried away from the lungs in blood, or lymph fluid that surrounds lung tissue. The survival rate of lung cancer persons decrease in the globe since the diagnosis of cancer cell not at right time hence the gradual increase in cancer growth rate leads to death. Lung cancer is due to the abnormal formation of cells in the lungs. These abnormal cells grow rapidly and divide to form tumour in the lungs. It is stated that the growth of these abnormal cells can spread beyond the Lungs and spread to other parts of the body (Parameshwarapa and Nandish, 2014). Lung cancer is diagnosed from the CT image of lung. The manual process of analysing the presence of cancer in lung may fail sometimes and it is not helpful to detect the cancer nodules accurately. Hence an automated and computerized method is needed for the detection of cancer nodules. Such automated and computerized system can be developed using image processing techniques to detect the lung cancer. Recently lot of image processing techniques are evolved and they are the best to detect the lung cancer nodules. Lung segmentation is an important pre-processing step occurring before nodule detection (Krishnamurthy et al., 2016) and the generation of a region of interest (ROI) for subsequent analysis (i.e the lung field).

(Lee et al., 2008) proposed a method which is based on the random forest leamer. The Training set contains nodule, non-nodule, and false-positive pattems. 5721 images selected from the LIDC lung databases. Test set contains randomly selected images. The proposed method is compared against the support vector machine. The proposed random forest based classifier performs well to detect all the nodules in the images and recorded a low false detection rate. It results 100% sensitivity and 1.27 FP/scan.

Shen et al., (2014) proposed a parameter-free lung segmentation algorithm with the aimof improving lung nodule detection accuracy,focusing on juxtapleural nodules. A bidirectional chain coding method combined with a supportvector machine (SVM) classifier is used to selectively smooth the lung border while minimizing the over-segmentation of adjacent regions. They tested this automated method on 233 computed tomography (CT) studies from the lung imaging database consortium (LIDC), representing 403 juxtapleural nodules. The results show that the method is able to correctly include the juxtapleural nodules into the lung tissue while minimizing over and under-segmentation. The limitation

*[1]MCA Department, St.Xavier's College, Tamilnadu,India, [2]Research Supervisor Dilla University, Ethiopia. *For Correspondence: ssararavi@yahoo.co.in, drsheela09@gmail.com*

of this method is that it sometimes fails to re-include the juxtapleural nodules sitting in consolidation regions (between lung tissue segments);

Ajil and Sreeram (2015), presented a novel method for lung nodule detection, segmentation and recognition using computed tomography (CT) images. In this work the lung area is segmented by active contour modeling followed by some masking techniques to transfer non-isolated nodules into isolated ones. Then, nodules are detected by the support vector machine (SVM) classifier using efficient 2D stochastic and 3D anatomical features. The proposed method is examined and compared with other efficient methods through experiments using clinical CT images and two groups of public datasets from Lung Image Database Consortium (LIDC) and ANODE09. Solid, non-solid and cavitary nodules are detected with an overall detection rate of 89%; the number of false positive is 7.3/scan and the locations of all detected nodules are recognized correctly.

Krishnamurthy et al., (2016), proposed an automatic three-dimensional segmentation algorithm which is used to segment the tissue clusters (nodules) inside the lung. However, an automatic morphological region-grow segmentation algorithm that was implemented to segment the well-circumscribed nodules present inside the lung did not segment the juxta-pleural nodule present on the inner surface of wall of the lung. A novel edge bridge and fill technique is proposed in this article to segment the juxta-pleural and pleural-tail nodules accurately. The algorithm proposed in this paper precisely detected 22 malignant nodules and failed to detect 3 with a sensitivity of 88%. Furthermore, this algorithm correctly eliminated 216 tissue clusters that were initially segmented as nodules; however, 41 non-malignant tissue clusters were detected as malignant nodules. Therefore, the false positive of this algorithm was 2.05 per patient.

Over the years, there are several different approaches and algorithms are developed for the automatic lung segmentation. But not all the methods are handled the juxtapleural nodules. Only a few methods include the juxtapleural nodules and sometime it fails to include the juxtapleural nodules sitting in consolidation regions ie between lung tissue segments. This paper handles the juxtapleural nodules efficiently using the proposed critical point selection algorithm which improves the computation time and the accuracy of the nodule deduction.

Materials and Methods

The three main steps involved in lung nodule detection process are pre-processing, feature extraction, and finally the classification process. The following block diagram (Figure 1) depicts the proposed work.

The proposed method consists of four steps 1) Pre-processing to generate lung boundary using adaptive thresholding and flood filling 2) Optimal critical points detection 3) Shape and size feature extraction for the optimal critical points and 4) Nodule detection classification with SVM and RF classifiers. The details for each step are described as follows.

Pre-Processing

Image pre-processing is the initial step for this proposed work. The main purpose of the pre- processing is to enhance the quality of the image and enhance the important image features and suppress the undesired ones. Pre processing is applied to the original CT scan image. Both lungs and their nearby portions are AOI (areas of interest) and pixel values external to this area being insignificant are removed. The pre processing of lung segmentation uses Otsu thresholding and other techniques of digital image processing. This method uses discriminate analysis to exhaustively search for a threshold value that minimizes the intra-class variance between two regions of an image. Otsu's thresholding chooses the threshold to minimize the intra class variance of the threshold black and white pixels. Threshold value is incremented step by step to reach a threshold value that gives maximum variance between pixels of the two classes. Let L represent the grey level (Shen et al., 2014) of all the pixels [1,2, …, L]. By choosing threshold at grey level k, the pixels are divided into object class C0 and background class C1. Let w0 and w1 be the probabilities of C0 and C1 separated by defined threshold and let σ_0^2 and σ_1^2 be the variances of these two classes. The intra-class variance is defined as the weighted sum of these two variances:

$$\sigma i^2_{ntra}(k) = \omega_0(k) * \sigma_0^2(k) + \omega_1(k) * \sigma_1^2(k) \quad (1)$$

The optimal threshold T is calculated as the value minimizing $\sigma^2_{intra}(k)$

$$T = \arg\min \sigma^2_{intra}(k) \quad (2)$$

After thresholding, flood fill algorithm is used to fill the lung area so that it can be further used for the process of segmentation (Mahale et al., 2017).

Optimal Nodule Criteria Boundary Pixels Detection

A lung boundary region is obtained after the pre-processing. In order to find the critical points in the boundary of lung lobe region we used bidirectional differential chain (BDC) encoding method. A chain code is a lossless compression algorithm for monochrome images. The basic principle of chain codes is to separately encode each connected component, or "blob", in the image. For each such region, a point on the boundary is selected and its coordinates are transmitted. The encoder then moves along the boundary of the region and, at each step, transmits a symbol representing the direction of this movement. The basic idea is to separately encode the boundary coordinates (chains of pixels) for each connected component in an image (Shen et al., 2014). The chain is a sequence of direction codes from one pixel to the adjacent one. There are eight possible (Gonzale, 2006) directions between two adjacent pixel and it is represented in figure 2 (b). To detect both horizontal and vertical critical points, this method uses two different coordinate systems for horizontal and vertical encoding. Horizontal code word generation results in horizontal code word and Vertical code word generation results in vertical code word. The encoder moves along the boundary following a (counter)

clockwise path, and at each step the direction of this movement is transformed into a code word as represented in Figure 2 (c).

A differential operation is used to generate the horizontal and vertical differential chain codes, separately. Non-zero points in the differential chain are identified as critical points. Figure 3 shows the detected critical points after the implementation of the bidirectional chain code. Figure3 a shows an original CT slice. The red circles on the boundary represent the horizontal critical points in fig. 3b and the green circles represent the horizontal critical points in Figure 3c.

Instead of processing area between each critical pair as described in the existing system (Shen et al., 2014), the proposed optimal critical point selection algorithm (OCPS) will select only those critical points (cp) which satisfy certain criteria. The OCPS algorithm defines two variables Minr and Maxr where Minr and Maxr represents the minimum and maximum distance range between two critical points. The value taken for Minr is 6 where as the value of Maxr is 20. After the first critical point selection the next critical point will be selected only if the distance between cpi and cpj is in between the range Minr and Maxr which reduces the number of critical points selected for the classification process, thereby reducing the computational complexity of the automated system. The following algorithm shows the steps involved in Optimal Critical Point Selection Algorithm.

Optimal Critical Point Selection Algorithm

Input: Bidirectional Chain Coding, Critical Points CP, Min_r, Max_r
Output: Optimal Critical points

Step1 : Initialize i=1; Fetch cpi €CP
Step2 : while i < NP
Step3 : j ← i+1
Step4 : $dist_{ij}$ ← distance(cp_i , cp_j)
Step5 : If $dist_{ij}$ >= Min_r and $dist_{ij}$ <= Max_r
Extract the features (Shape & Size Features)
Classify if nodule or non nodule using classifiers (SVM & RF)
i ← j + 1
Go to step 2.
Else
j ← j + 1
Go to step 4

The following figure 4 shows the extracted horizontal and vertical critical points using optimal critical point selection algorithm for a single slice for the patient id 11 in the LIDC dataset. In this case the total number of critical points extracted in the existing system is 113 where as in the proposed OCPS is 35.

Feature Extraction

After extracting the region of interest, we have to find whether all the extracted regions are nodules or not. Analysis should be done on the features of each nodule to distinguish true nodules from the false positive nodules. Many features, such as size of area, circularity, mean value, variance, location, or gradient, etc, are usually applied in detection of medical images (Varalakshmi, 2013). The feature extraction process is very important stage in image processing technique that uses algorithms and techniques to detect and isolate various desired portions or shapes (features) of an image (prasad, 2013). Feature extraction is an essential stage that represents the final results to determine the normality or abnormality of an image (kumar et al., 2016).

Shape Features

In order to quantify the shape of a nodule, we used some common image shape features: concave, length, position, circularity, eccentricity, solidity, extent and Euler number (Raicu et al., 2007).

Size Features

For the size of a nodule, the following five size features were found to be the most common ones: area, convexArea, perimeter, majorAxisLength and minorAxisLength.

Figure 5 shows the values for shape and size features extracted for a segmented nodule.

Classification

The final procedure of the proposed system is to confirm the suspicious region and determine if it is a true nodule utilizing features obtained from previous stages. In this paper the classification process is done by using Support Vector Machine (SVM) and Random Forest (RF) Classifiers. Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification (Tidke and Chakkarwar, 2012). The basic SVM takes a set of input data and for each given input, predicts which of two classes forms the input, making it a non probabilistic binary linear classifier. With SVM each data item is plotted as point in n dimensional space where n is number of features (Kulkarni and Bagal, 2016). The main objective of SVM is to find the hyper plane that gives largest minimum distance to training example. Various types of kernels such as Gaussian, Radial basis kernel, Linear, polynomial can be implemented in SVM (Ajil and Sreeram, 2015). The Random Forest algorithm is one of the best among the classification algorithms. It is able to classify large amounts of data with accuracy. Random Forests are an ensemble learning method (also thought of as a form of nearest neighbour predictor) for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees (Lee et al, 2008). Random Forests are a wonderful tool for making predictions considering they do not over fit because of the law of large numbers. Introducing the right kind of randomness makes them accurate classifiers.

Results

About Data Set

CT images from various stages of lung cancer are needed to carry out the classification process. The images are strictly in DICOM (Digital Imaging and Communication in Medicine) format in order to maintain medical standardization. The lung CT images are obtained from Lung Image Data Consortium (LIDC). The proposed method was validated using data from LIDC, available

Table 1. Performance Comparison of SVM and RF

Methods	Precision	Recall	Fscore	Accuracy
BDC_SVM	1	0.523179	0.686957	0.641791
BDC_RF	1	0.596026	0.746888	0.696517
OCPS_SVM	1	0.655629	0.792	0.741294
OCPS_RF	1	0.748344	0.856061	0.810945

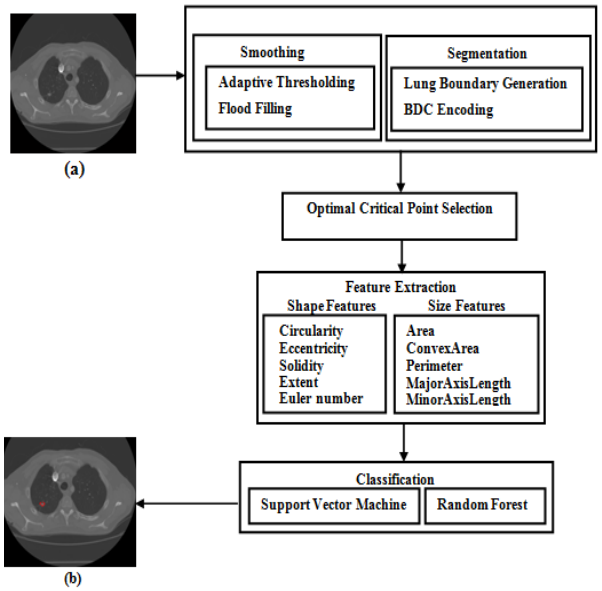

Figure.1 Diagram Describing the Proposed Method a) Input Image b) Output Image

through The Cancer Imaging Archive (TCIA). LIDC comprises thoracic imaging studies gathered from five sites across the United States (Armato and McLennan, 2011). The LIDC database has lung CT scans of 10 patients with the ground truth information about nodules position in separate excel sheet.

Evaluation Method

To evaluate the performance of the proposed method several performance metrics are available. This paper uses the Precision Rate, Recall Rate and F-Measure to examine the performance.

Precision Rate

Precision is calculated as the fraction of correct points among those that the algorithm believes belonging to the relevant class. It can be loosely equated to accuracy.

Precision= TP/(TP+FP)

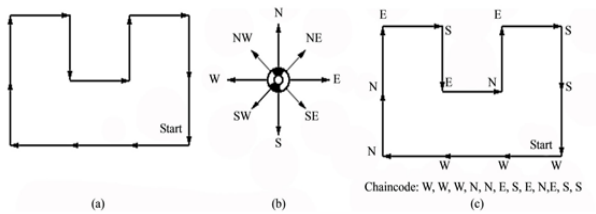

Figure 2 Bidirectional Chain Code Representations

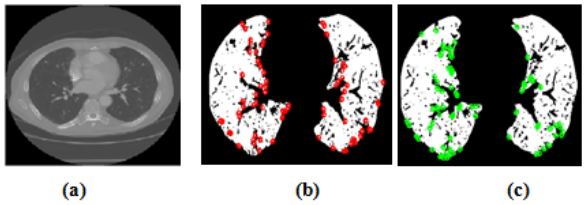

Figure 3. (a) Input Image (b) Deducted Horizontal Critical Points (c) Deducted Vertical Critical Points

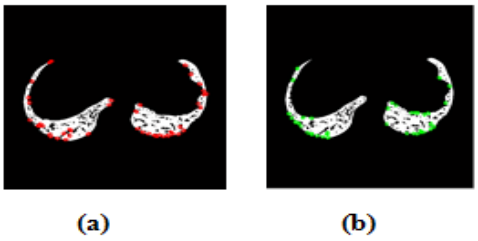

Figure 4. (a) Horizontal Critical Points Using OCPS, (b) Vertical Critical Points Using OCPS

Where TP = True Positive (Equivalent with Hits)
FP = False Positive (Equivalent with False Alarm)

Recall Rate

Recall is the fraction of relevant instances that are

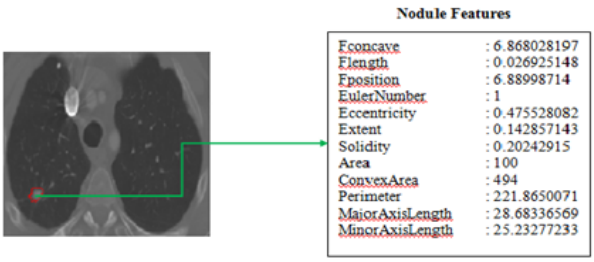

Figure 5 Features Extracted from the Segmented Nodule

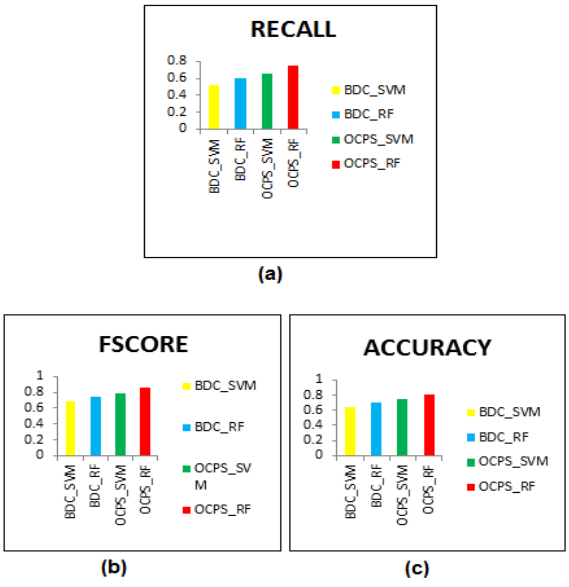

Figure 6. (a) Performance Analysis for Recall Rate (b) Performance Analysis for F-Score (c) Performance Analysis for Accuracy.

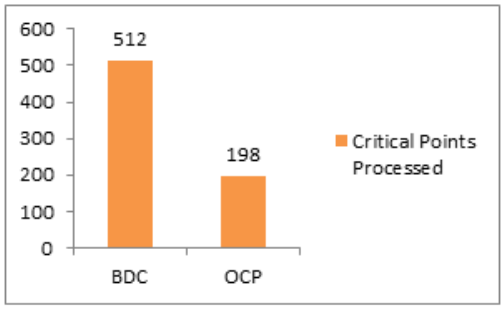

Figure 7. Optimal Critical Points Processing Analysis

retrieved.

Recall= TP/(TP+FN)

Where TP = True Positive (Equivalent with Hits)
FP = False Negative (Equivalent with Miss)

F-Score

F-measure is the ratio of product of precision and recall to the sum of precision and recall. The f-measure can be calculated as,

F_m=(1+α)* (Precision*Recall)/(α*(Precision*Recall))

Accuracy

Accuracy is how close a measured value is to the actual (true) value.

Accuracy= (TP+TN)/(TP+TN+FP+FN)

The following tabular column table I shows the performance comparison between the BDC and the proposed approach OCPS with SVM and RF classification algorithms in terms of Precision Rate, Recall Rate, F-Measure and Accuracy. Figure 6 shows the performance analysis for Precision Rate, Recall Rate, F-Measure and Accuracy.

Analysis

The proposed system uses LIDC database of lung cancer patients. CT images in DICOM format are successfully undergone through pre-processing. The proposed Optimal Critical Point selection algorithm is used to deduct the critical points. The shape features namely concave, length, position, circularity, eccentricity, solidity, extent and size features such as area, convexArea, perimeter, majorAxisLength, minorAxisLength are extracted for the critical points. These extracted features are stored in excel sheet and provided further to SVM and RF classifier for further classification. The SVM classifier and random forest classifier are trained with the nodule and non-nodule patterns. The random forest classifier and support vector machine classifier were tested with lung CT images chosen from the LIDC database. The codes for training and testing both the classifier based system were written and executed in MATLAB. The following graph shows the total number of computations for the patient ID 11from LIDC data set which contains 10 slices. Bi directional chain code computed 512 critical points where as the proposed optimal critical point selection (OCPS) computed only 198 critical points. The computation time taken by optimal critical point selection algorithm along with classification for the patient ID 11 is 17.804114 and the computation time needed for the existing method is 29.6348012 seconds which indicates that the optimal critical point selection algorithm reduces the computational complexity.

Discussion

The proposed work aims to find a better segmentation algorithm which focuses on the accuracy of juxtapleural nodule detection. The proposed automatic detection of lung nodule system consists of three stages that are image pre processing, nodule detection and classification. Lung CT Images are selected from the LIDC lung databases. Lung CT Images are pre processed using Otsu thresholding and flood filling method. The nodule criteria region extraction is done by using the proposed Optimal Critical Point Selection Algorithm to detect the juxtapleural nodule. Shape and size features are calculated for the segmented nodules. Support Vector Machine (SVM) and Random Forest (RF) are used as classifiers to implement the classification process. This automated method was tested on lung computed tomography (CT) studies from the lung imaging database consortium (LIDC), representing 406 juxtapleural nodules. From the results it is observed that the optimal critical point selection (OCPS) algorithm processed only 282 juxtapleural nodules and reduces the computational complexity. In the classification process the Random Forest based classifier method performs well to detect all the nodules in the images which provides improved accuracy and reduces the false positive rate than the Support Vector Machine. We conclude that the combination of Random Forest Classifier along with Optimal Critical Point Selection algorithm performs well and increases the accuracy of the automated lung cancer detection process. This work can be extended in future by analyzing the texture feature of the juxtapleural nodules along with the classifiers.

References

Ajil MV, Sreeram S (2015). Lung cancer detection from CT image using image processing techniques. *IJARCSMS*, **3**, 249 – 5.

Armato SG, McLennan G, Bidaut L, et al (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med Phys*, **38**, 915–31.

Gonzale Z, Woods E (2006). Digital image processing, representation and descripstion, prentice hall, upper saddle river, New Jersey, pp 644 – 9.

Krishnamurthy, Narasimhan, Rengasamy (2016). Three-dimensional lung nodule segmentation and shape variance analysis to detect lung cancer with reduced false positives. *Proc IME H J Eng Med*, **230**, 58–70.

Kulkarni G, Bagal B (2016), Lung cancer tumor detection using image processing and soft computing techniques, *IJSTM*, **5**, 451 - 8.

Kumar S, Swathy M, Sathish S, Sivaraman J, Rajasekar M (2016). Identification of lung cancer cell using watershed

segmentation on CT images. *INDJST*, **9**, 1- 4.
Lee SLA, Kouzani AZ, Hu EJ (2008). A random forest for lung nodule identification. *IEEE*, pp 1-5.
Mahale A, Rawool C, Tolani D, Bathija D, Jewani K (2017). A survey on lung cancer detection using image data analysis and machine learning. *Int J Innovat Res Comput Comm Eng*, **5**, 1066- 75.
Parameshwarapa V, Nandish S (2014), Segmentation of lung cancer using image enhancement techniques and region growing algorithm. *Int J Eng Res Tech*, **3**, 482- 85.
Prasad (2013). Lung cancer detection using image processing techniques. *Int J Latest Trends Eng Tech*, **3**, 372– 8.
Raicu S, Varutbangkul E, Cisneros G, et al (2007). Semantics and image content integration for pulmonary nodule interpretation in thoracic computed tomography. *Medical Imaging*, **6512**, 1- 12.
Shoaib M, Naseem R, Dar H (2013). Automated segmentation of lungs in computed tomographic images. *Eur J Sci Res*, **98**, 45 - 54.
Shen S, Bui AT, Cong J, William Hsu (2014). An automated lung segmentation approach using bidirectional chain codes to improve nodule detection accuracy. *Comput Biol Med*, **57**, 139–49.
Tidke P, Vrishali A (2012). Classification of lung tumor using SVM. *Int J Comput Eng Res*, **2**, 1254 - 7.
Varalakshmi K (2013). Classification of lung cancer nodules using a hybrid approach. *J Emerg Trends Comput Inform Sci*, **4**, 63- 8.