A Novel Diagnostic Model For Lung Cancer Detection using Mumford-Shah and SVM Classifier

Kishore Sebastian, S. Devi

Abstract*: Lung cancer is one of the very deadly diseases in the world. However, diagnosing it at an early stage and treating it properly can protect lives. Although Computer Tomography (CT) scan imaging is one of the fruitful imaging in the field of medicine, it is the hardest for clinicians to clarify and recognize cancer from those images. And it is carried out with Mumford and Shah functional model, and support vector machine (SVM) classifier. Also, the system takes less computation time and thus, is highly efficient than existing algorithms which grab 98% accuracy. Further, the performance analysis of the proposed system is executed using seven assessment metrics namely Classification Accuracy, sensitivity, specificity, AUC, F measure, Precision, Brier Score, MCC. Finally, the results of SVM are compared with KNN, Decision-Tree, and Adaptive Boosting algorithms in terms of the seven metrics. The results show that there is significant progress in the above measures than the existing method.

***Keywords* : Lung Cancer, CT Scan, Mumford-Shah Model**

I. INTRODUCTION

Lung cancer is one of the leading reasons of cancer death. It is hard to find, since it grows and it show signs only at the final stage. However, detecting it at an earlier stage and treating the disease can reduce mortality and probability. The fruitful imaging technique is CT imaging. It is credible for the detection of lung cancer, as it may expose every lung cancer noodles that is suspected [1]. However, the variability of the intensity in images and the anatomical misalignment by physicians and radiologists may produce dilemma in labeling the cancer nodule [2]. Recently, computer-assisted diagnostic has become a promising and reliable tool for radiologists and clinicians to accurately diagnose cancer [3]. Several research methods are being developed to detect lung cancer. However, some of the research works does not show any satisfactory accuracy in the detection of the disease and the works still need to be improved to grab high accuracy closer to 100%. Image processing methods [10] and machine learning methods have been implemented in order to detect and also to classify the type of lung cancer. The latest systems so far developed for lung cancer detection based on CT scan images are thoroughly reviewed, analyzed and proposed a new model.

Revised Manuscript Received on January 05, 2020

Kishore Sebastian*, Department of Computer Science and Engineering, PRIST University, Thanjavur, India.

S. Devi, Department of Electronics and Communication Engineering, PRIST University, Thanjavur, India.

II. LITERATURE REVIEW

Many research works have been done on the diagnosis of lung cancer using various approaches to image processing and also machine learning. The authors [4] proposed a model that classifies between knots and normal anatomy of the lung. The model extracts the properties such as geometrical, statistical, and gray level features and classifies using LDA and as a gateway to separation. The model achieves accuracy of 84%, sensitivity of 97.14% and specificity of 53.33%. Although the model detects cancerous nodule, its accuracy is low and also the model uses no machine learning techniques and simple classification techniques are used. The authors [5] used the neural network for classification in CAD system to detect lung cancer. Their system achieved around 85% accuracy, sensitivity and specification. Their work promises to reduce the cost of implementation but still its accuracy is unsatisfactory. The K-means algorithm for clustering is used to compile the pixel dataset according to some features. This model performs classification using Neural network and the features such as entropy, correlation, integrity, BSNR, and SSIM are extracted using the gray-state co-occurrence matrix (GLCM) to achieve 90.7% accuracy [6]. The authors [7] developed a method for diagnosing lung cancer using an ambiguous interference system using gray conversion and binarization and an active margin model for decomposing the result. Classification of cancer cells is performed using the fuzzy inference system (FIS) which utilizes the extracted features for its training. Their model achieves the accuracy of 94.12%. The authors [8] developed a system using watershed method for segmentation and uses Gabor filter for pre-processing and achieves the accuracy around 90%. The review extensively showed that the above works does not differentiate the cancer cells as benign or malignant. The system [9] differentiate s lung cancer as benign or malignant using the primary data and the Housefield Unit (HU) in order to calculate the region of interest (ROI). The circularity, posterior dimension, area, eccentricity and shape features are extracted to train the SVM and also to classify in order to identify the benign or malignant nodule.

Kishore Sebastian and S. Devi [10] developed a system based on Mumford and Shah model for diagnosing lung cancer by analyzing CT images of the lung. The system

includes filtering, image segmentation, binarization and thinning, minimization of false minute points and finally the GLCM is for feature extraction. Moreover, the system eliminates

98% of the noises, making it more efficient than other methods, reaching 80% to 90%. The work does not include the classification of ROI using the available types of classifiers. The proposed system uses the same method [10] and classifies cancer nodule as benign or malignant using a SVM. Also the results obtained using SVM classifier is compared with KNN, Decision-Tree, and Adaptive Boosting algorithms.

III. PROPOSED MODEL

The proposed work in this study aims to design a good CAD method for detecting lung cancer. It has four phases. They include the extraction of the lung area from chest CT images, the segmentation of the lung, the feature extraction, and the classification of cancer cells as benign and malignant nodules. The second stage is implemented using the Mumford and Shaw model, the third stage is implemented using GLCM and the fourth stage is performed using SVM, KNN, end-tree and adaptive boosting algorithms and their performance is compared. Experiments are performed on real-time lung imaging and LIDC-ITRI-assisted computer-assisted diagnostic systems [13]. Almost 500 real-time images of different types of people were collected for the study. These films have benign and malignant nodules. In these images, 270 patients with no signs of cancer are discharged, and the remaining 230 patients are analyzed for benign and malignant tumors. Malignant nodules are not cancerous, benign nodules are not cancerous. Images are 1 mm from the slice thickness scanner, the voltage peak is 120 kVP, the tube current is 220 mA, and the scan rate is 300 pieces per minute. Each image is 512 x 512 pixels in size as an unsigned integer. Modifications have been made to the existing solution [10], and the proposed model is shown in Fig. 1. The pre-processing phase uses sigma filter and after the pre-processing phase, the processed image is segmented using the Mumford and Shah functional model. It provides image with significant cancer nodule. In addition, the features such as area, circumference, and eccentricity, features such as centroid, diameter, and pixel mean intensity have been extracted during the feature extraction phase for detected cancer nodule. The model is efficiently improved only after the cancer is diagnosed, which is feature extraction and calculation of accuracy. But its classification is not implemented as benign or malignant. Therefore, an additional phase has been made to classify cancerous knots using the SVM approach using thee extracted features as training features and a trained model is created. Then, the unknown diagnosed cancer knot is classified using the trained predictive system. Figure 1 shows the module design for the proposed method.

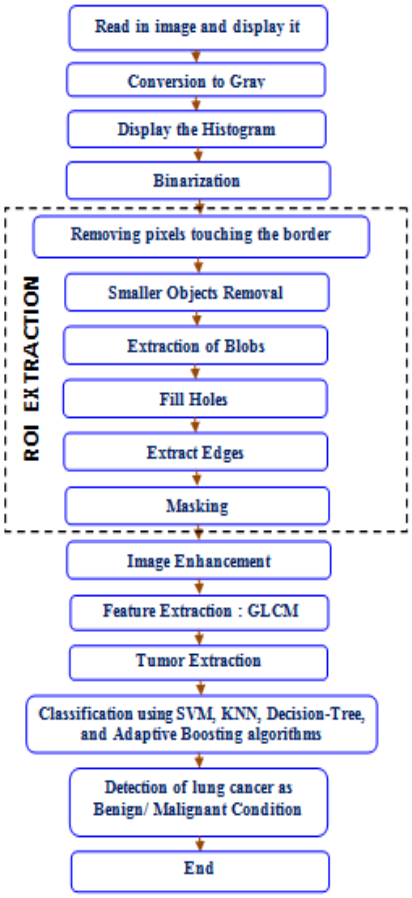


Fig.1 Module Design for the proposed method

A. Conversion To Gray

The luminance pixel of a grayscale image ranges from 0-255. Converting a color image to a grayscale image converts RGB values of 24 bit to a grayscale value of 8 bit. This section transform color into grayscale image.

B. Histogram

The threshold value (T) is chosen that correctly separates objects (light) from the dark background. Image histograms are the means to view the distribution of grayscale intensity values across the entire image.

C. Binarization

Image binarization takes a grayscale image and make it to black and white, basically reducing the image information from 256 shades of gray to two: black and white. Here an image is segmented into constituent objects.

D. ROI Extraction

To extract the ROI of any domain, the Mumford–Shah functional Model is used. Typically, the ROI of the segment includes edge detection, thresholding and clustering, or detection of edges from an image [11].

E. Removing Pixels Touching the Border

Next to segmentation, the image has objects usually that touch the boundaries of the image. It is best to remove the objects since it is not possible to get the complete information about them.

F. Smaller Objects Removal

This section removes all connected elements (objects) that are less than P pixels from the binary image, creating another binary image

G. Extract Only the Largest Blobs

Blob refers to the binary large object and the group of pixels enclosed in the binary image. The word " Large " refers to the fact that only a certain number of objects are interested, and "small" binary objects are generally noisy. The purpose of Blob extraction is to isolate objects of interest in a binary image.

H. Fill Holes

Often, after some morphological operation, it is needed to fill holes in a binary image.

I. Extract Edges.

Canny edge detector is a multi phase algorithm for edge detection. The multi phase includes Preprocessing, Calculation of gradients, Non-maximum suppression, Thresholding with hysterysis.

J. Masking

A mask consists zeroes and non-zero values. It is also a binary image and if it is applied to another binary/ grayscale image of similar size, all the pixels become zero in the resultant image if there is a zero in the mask while the other pixels remains same.

K. Image Enlargement

Image enhancement adjusts digital images so that the results are more suitable for visualization or further image analysis. Adaptive Histogram Equilibrium (AHE) utilizes the transformation operation (function derived from the neighborhood) in order to convert all the pixels of the image. That function is directly proportional to the distribution function of the pixel values in the neighbor region.

L. GLCM feature extraction

The Gray Level Co-Occurrence Matrix (GLCM) is a popular statistical system for extracting textual features. The GLCCM defines fourteen such features that are determined from the probability matrix for extracting the properties of the image texture statistics.

M. Tumor extraction

To separate the nodules, the CT images should be divided into two parts. One area contains nodules cells, and the second part contains normal cells.

N. Classification

In this section, the condition is classified as malignant or benign using SVM which is a supervised machine learning system. A support vector machine (SVM) is a discriminant classifier, formally defined by a dividing hyperplane. In other words, given the labeled training data (supervised learning), the algorithm releases an optimal hyperplane, which classifies new data. This hyperplane in two-dimensional space is a line dividing a plane into two sections, where the classes lies on two sides.

IV. PERFORMANCE MEASURES

Image quality is the characteristic of an image, which is used to measure the processed image by comparing it to an ideal image. In this study, various image quality metrics are considered and their statistical behavior are analyzed. The various assessment metrics include F-Measure, Area under ROC curve (AUC), Accuracy, Specificity, Matthews Correlation Coefficient, and Brier Score respectively. Table 2 shows the various assessment metrics in the analysis of image classification. From the Table 1, it is found that F-Measure, Area under ROC curve (AUC), Accuracy, Specificity, Matthews Correlation Coefficient, and Brier Score are increased than the existing model.

A. F-Measure

F-measure measures the performance and success of the segment based on precision (p) and recall (r) values. To have a measurement with high efficiency, the single, F measurement is calculated with precision and recall. This is a compatible average, which is defined using the formula in equation 1 to give an accurate result.

$$F\text{-}Measure = 2/((1/p)+(1/r)) \quad (1)$$

B. Area under ROC curve (AUC)

This is a simple measurement metric that can be used to measure accuracy by reducing the ROC curve as a measurable value. The value is normalized between 0 and 1 range. The higher value of the AUC indicates the better performance of the segment. It is calculated using the formula given below in equation 2,

$$\text{Area under ROC curve (AUC)} = \int_g^t x(a) \quad (2)$$

Let g and f be the top and bottom of a function curve with the minimum and maximum axis points of the x (a) curve.

C. Accuracy

Accuracy is the percentage of instances that are correctly classified. It is given as (TP + TN) / (TP + TN + FP + FN). Where TP, FN, FP and TN are the number of true positives, false negatives, false positives, and true negatives, respectively. For good classifiers, both true positive rate (TPR) and true negative rate (TNR) must be close to 100%.

D. Sensitivity

The sensitivity/ TPR is defined as the total number of pixels correctly detected as abnormal cells. It is given by TP/(TP+FN).

E. Specificity

Specificity/ TNR is the percentage of pixels that are identified as normal, albeit abnormal. It is given by FP/(FP+TN).

F. Matthews Correlation Coefficient.

The Mathews correlation coefficient (MCC) has a range of -1 to 1, where -1 represents a completely invalid binary

classifier, and 1 represents a completely valid binary classifier. Using MCC allows them to know how well their classification model / function works. MCC is given by the

equation 3.

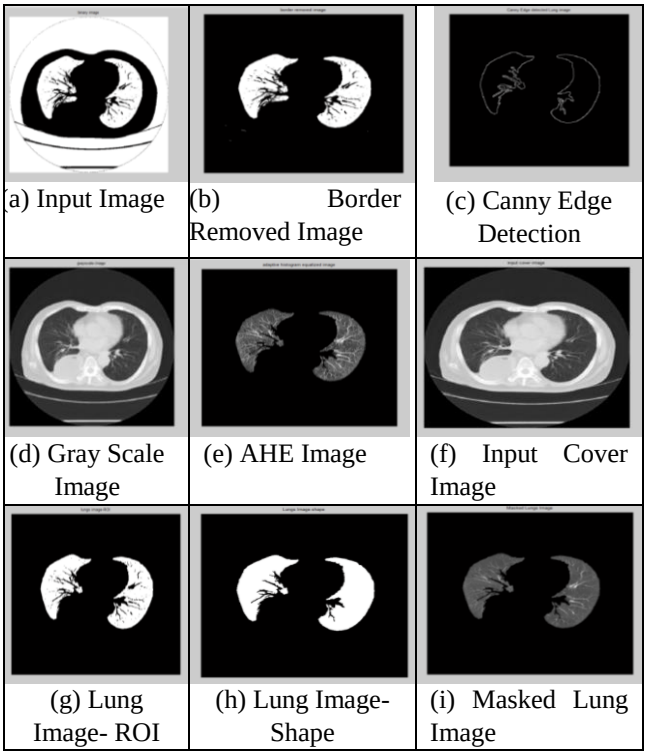
$$\text{Mathews correlation coefficient (MCC)} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3)$$

G. Brier Score

The Brier score, computes the mean squared error between the predicted probabilities and the expected values. The score summarizes the degree of error in probability predictions. The error score will always be between 0.0 and 1.0, where the model with the correct skill will score 0.0.

V. SIMULATION RESULTS

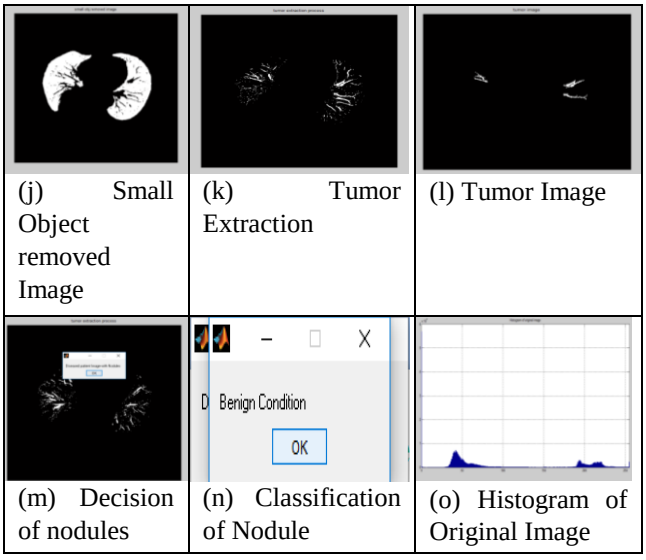
Simulation of the proposed system is performed in MATLAB [13]. And the results are shown in Figure 2. Figure 2(a) shows the sample image taken as input, Figure 2(b) shows Border Removed Image, Figure 2(c) shows Canny Edge Detection, Figure 2(d) shows Gray Scale Image, Figure 2(e) shows AHE Image, Figure 2(f) shows Input Cover Image, Figure 2(g) shows Lung Image- ROI, Figure 2(h) shows Lung Image- Shape, Figure 2(i) shows Masked Lung Image, Figure 2(j) shows Small Object removed Image, Figure 2(k) shows Tumor Extraction, Figure 2(l) shows Tumor Image, Figure 2(m) shows Decision of nodules, Figure 2(n) shows Classification of Nodules Figure 2(o) shows Histogram of Original Image respectively. The major advantage of the proposed model are there is an increase in the accuracy of cancer endpoint detection than the existing model. The results shown a progressive improvement in accuracy from 88.4% to 98%. Also the diagnosis of lung cancer is classified as malignant or benign.



(a) Input Image | (b) Border Removed Image | (c) Canny Edge Detection

(d) Gray Scale Image | (e) AHE Image | (f) Input Cover Image

(g) Lung Image- ROI | (h) Lung Image- Shape | (i) Masked Lung Image



(j) Small Object removed Image | (k) Tumor Extraction | (l) Tumor Image

(m) Decision of nodules | (n) Classification of Nodule | (o) Histogram of Original Image

Figure 2. Simulated outputs of the proposed system

Table I shows the comparison of the existing edge detection method (Sobel) and the proposed edge detection method (Canny) in terms of Accuracy and Time. It is found that Canny edge detection has more accuracy compared to Sobel and the computation time is also lesser than Sobel.

Table I Canny and Sobel

Algorithms	Accuracy (%)	Time (Sec)
Sobel	76%	32
Canny	91.2%	18

Table II shows the comparison of all the classifiers namely SVM, Adaboost, KNN, Neural Network [12], and Decision Tree in terms of Classification Accuracy, sensitivity, specificity, AUC, F measure, Precision, Brier Score, Matthews Correlation Coefficient.

Table II. Performance of Classifiers

Method	η	μ	σ	AUC	F1	υ	ρ	£
SVM	0.994	1.000	0.987	0.999	0.994	0.987	0.011	0.987
Adaboost	0.987	0.987	0.988	0.987	0.987	0.987	0.026	0.975
KNN	0.981	0.974	0.988	0.978	0.980	0.961	0.366	0.962
Neural Network	0.979	0.972	0.979	0.972	0.976	0.961	0.257	0.958
Decision Tree	0.962	0.961	0.963	0.986	0.961	0.961	0.072	0.924

η - Classification Accuracy, μ - sensitivity, σ - specificity, AUC- area under the curve, F1-F measure, υ - Precision, ρ -Brier Score, £- Matthews Correlation Coefficient

Figure 3 shows the accuracy performance of the classifiers such as SVM, Adaboost, KNN, Neural Network, and Decision Tree.

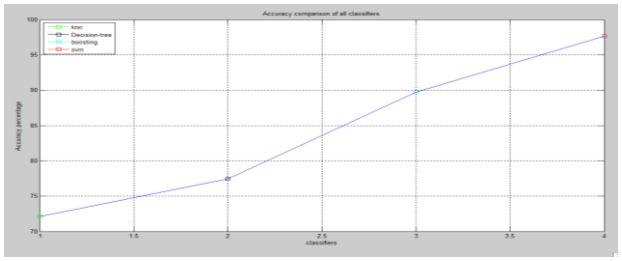


Fig. 3 Accuracy Performance of Classifiers

VI. RESULT AND EVALUATION OF IMPLEMENTATION

The results of all classifiers are tabulated in Table II. TP, TN, FP, FN nodes detected by the system are recorded. The proposed system achieves the accuracy of 98%. The extracted features such as area, circumference, centroid, diameter, eccentricity, and mean intensity of pixels were used to train the SVM method and a trained model is created. The training time for the classification is 4.3 seconds which is an improvement over the existing method [10].

VII. CONCLUSION

In this paper, the proposed method uses SVM classification with Mumford and Shaw to detect lung nodule, thereby improving the effectiveness and robustness of the system. Therefore, the system achieves 98% accuracy, which makes it more efficient than other methods. This system may mark the stage of lung cancer at an earlier stage. Therefore, the function of this system plays an important role in preventing aggressive phases and reducing the percentage of lung cancer in mankind. With the support of additional information for radiologists and clinicians, the system helps to make the right decision about lung cancer with high accuracy at the right time.

REFERENCES

1. Gindi,A. M, Al Attiatalla, T. A, and Sami, M.M, "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." Journal of American Science, vol. 10, 2014, pp. 13-22.
2. Suzuki. K, Kusumoto. M, Watanabe. S. I, Tsuchiya. R., and Asamura. H. (2006) "Radiologic classification of small adenocarcinoma of the lung: radiologic-pathologic correlation and its prognostic impact," The Annals of Thoracic Surgery, vol. 81, 2006, pp. 413-419.
3. Xiuhua,G, Tao, S, and Zhigang. L, "Prediction Models for Malignant Pulmonary Nodules Based-on Texture Features of CT Image." In Theory and Applications of CT Imaging and Analysis, 2011, pp. 63-76.
4. Aggarwal, T, Furqan, A, and Kalra. K, "Feature extraction and LDA based classification of lung nodules in chest CT scan images." 2015 International Conference On Advances In Computing, Communications And Informatics (ICACCI), 2015, Kerala, India.
5. Jin, X., Zhang, Y., and Jin, Q. "Pulmonary Nodule Detection Based on CT Images Using Convolution Neural Network." 2016 9Th International Symposium On Computational Intelligence And Design (ISCID), 2016, Hangzhou, China.
6. Sangamithraa, P., & Govindaraju, S, "Lung tumour detection and classification using EK-Mean clustering." 2016 International Conference On Wireless Communications, Signal Processing And Networking (Wispnet), 2016, Chennai, India.
7. Roy, T., Sirohi, N., and Patle, A, "Classification of lung image and nodule detection using fuzzy inference system." International Conference On Computing, Communication & Automation, 2015, Noida.
8. Ignatious, S., and Joseph, R. "Computer aided lung cancer detection system." Global Conference On Communication Technologies (GCCT), 2015, Thuckalay ,Nagercoil, India.
9. Rendon-Gonzalez, E., & Ponomaryov, V. (2016) "Automatic Lung nodule segmentation and classification in CT images based on SVM." 9Th International Kharkiv Symposium On Physics And Engineering Of Microwaves, Millimeter And Submillimeter Waves (MSMW)I, 2016, Kharkiv, Ukraine.
10. Kishore Sebastian & S. Devi, "A novel model of feature extraction for lung cysts detection in CT image using Minutiae based Mumford and Shah functional model", Australian Journal of Electrical and Electronics Engineering, vol. 16, September 2019.
11. Miah, M.B.A., and Yousuf, M.A, "Detection of lung cancer from CT image using image processing and neural network." International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2015, Dhaka, pp.1-6.
12. Khobragade, S., Tiwari, A., Patil, C., and Narke, V. (2016) "Automatic detection of major lung diseases using Chest Radiographs and classification by feed-forward artificial neural network." IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), 2016, Delhi, India, pp. 1-5.
13. Armato, I., Samuel McLennan, G., McNitt-Gray, F. R., Michael, Charles, Reeves, Anthony P., … Clarke, Laurenc, "Data From LIDC-IDRI. The Cancer Imaging" Archive.http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX.

AUTHORS PROFILE

Kishore Sebastian, Prof. completed B.E. in 2005 and M.E in 2007. His area of interest is Image Processing. He has published many journals. Presentlly he is working in St.Joseph's College of Engineering and Technology, Kerala.

Dr. S. Devi, completed B. E. in Electronics and Communication Engineering in Bharathidasan University, Trichy and M.Tech in Communication Systems in Dr. M.G.R University, and Ph.D in Anna University. She has more than ninteen years of experience. She has published more than twenty papers in reputed peer reviewed Journals, Conferences, and book chapters.