# A New Automated Approach for Early Lung Cancer Detection with Improved Diagnostic Performance – A Preliminary Analysis

**Thavavel Vaiyapuri1**

*Assistant Professor, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia*

*Abstract*—Lung cancer is becoming the major cause of cancer-related deaths in human worldwide and Saudi Arabia is not an exception. Therein the identification of potentially malignant lung nodules is essential for the diagnosis and clinical management of lung cancer. Unfortunately, in clinical practice, however, interpretation of Computed Tomography (CT) images is challenging for radiologists due to the large number of cases. It is therefore extremely important task to develop computer aided diagnosis (CAD) systems that can aid and enhance the radiologist to potentially reduce false positive (FP) findings. Even though numerous methodologies are proposed for CAD system in the literatures, the one proposed in this work will definitely stand out in improving the sensitivity and specificity for the detection of small nodules particularly in low dose CT images. This work attempts to employ the powerful tool, radiomics quantitative imaging features within curvelet domain to detect and characterize lung nodules with improved sensitivity and specificity. Subsequently, support vector machine (SVM) is used to learn the 2D stochastic and 3D anatomic features of curvelet coefficients and classifies suspected regions either as malignant or benign based on geometric, texture and intensity. A preliminary analysis of the proposed methodology is presented and compared against the metrics, sensitivity, specificity and accuracy on publicly available LIDC database to serve as a benchmark for future research efforts.

***Keywords:*** *Medical image Processing, Low-Dose Computed Tomography, Lung Cancer Detection, Computer-Aided Detection, Anisotropic Diffusion, Radiomic Image Features, Curvelet Transform and Support vector machine*

## Introduction

Lung cancer stands out among all other types of cancer with highest incidence rates and highest mortality rates in human worldwide[1-2]. The prevalence is steadily escalating and alarming due to adoption of western lifestyle habits[3]. Also, it is expected that in coming years lung cancer will represent a major public health problems[4]. In practice, Lung cancers are reported at advanced stages and this dramatically diminishes the potential success of treatment and results in the death of around 50% of cancer patients within five years of the late diagnosis[2]. If lung cancer is detected at its earliest stage, the five-year survival rate can reach 70%[3]. These figures call for effective technique to detect lung cancer at early stage to reduce the mortality rate of the patient and improve their life expectancy[21].

In lung cancer research, CT has become the most sensitive imaging modality for the detection of pulmonary nodules, particularly since the introduction of helical multi-slice technology. In current clinical practice, Low-Dose CT is adopted for Lung Cancer screening to avoid the risks of ionizing radiation associated with CT[5]. Unfortunately, the interpretation of Low-Dose CT images is challenging for radiologists to distinguish between benign and malignant nodules. Added to, the large number of cases and manual reading may cause error-prone and the reader may miss nodules and thus a potential cancer. Thus, there is a pressing need for the use of CAD systems that can aid/enhance radiologist workflow and potentially improve nodules detection.

Current CAD systems often only allow radiologists to visually describe the tumors or lesions[6] and thus are limited to subjective and qualitative characterizations of the tumors. This is because the algorithms employed in current CAD have high sensitivity that some non-nodule structures (e.g., blood vessels) are labelled as nodules inevitably in the initial nodule identification step. Since the radiologists must examine each identified object, it is highly desirable to eliminate these FPs as much as possible while retaining the true positives (TPs). Thus, it is very time-consuming for radiologists to perform extensive review of all available imaging data. As such, there is significant potential for improving diagnostic accuracy and efficiency through the use of more objective and quantitative approaches for tumor characterization.

One of the biggest emerging areas in recent years related to quantitative cancer screening and diagnosis is radiomics[7], which involves the high-throughput extraction and analysis of a large number of imaging-based features for quantitative characterization and analysis of tumor. The use of radiomics-driven approaches allows for a more objective and quantitative evaluation and diagnosis of cancer, which can significantly reduce inter-observer and intra-observer variability and improve diagnostic accuracy and efficiency compared to current qualitative cancer assessment strategies. A number of studies[8] have shown that radiomics can be used to characterize tumors and can have clinical significance towards diagnosis.

Despite its potential for huge clinical impact, the current state-of-the-art radiomics techniques make use of pre- defined, hand-engineered imaging-based feature models based on texture, shape, and intensity, which can limit its potential for fully characterizing the unique traits and characteristics of lung cancer lesions as they are still largely based on specific visual traits that radiologists use for subjective interpretation[9]. Therefore, an effective way that can identify and extract imaging features automatically compared to pre-defined and hand-engineered imaging feature models is expected to have potential in capturing tumor phenotype and improve lung cancer detection[10]. Recently, Shen et al[11] and Kumar et al[12] have shown the feasibility of discovering radiomics using a deep convolutional neural network learning rather than using predefined, hand engineered feature models. The main disadvantage of neural network methods lies in training process to adjust the network weights by supervised learning with standard training data. Secondly, the model performance is based on the large dataset employed for training process. Not only that, for each image in the training set, nodule structure needs have to be precisely marked by a radiologist is mandatory process.

In this essence, the proposed work first investigates the association between radiomic features based on curvelet trans- form and lung tumor characterization. Next it employs machine-learning method, SVM to build radiomics-based multivariate classifiers for tumor characterization. Non-invasive and cost-effective radiomic data is expected to improve the histological classification and hence the treatment/therapy, which in general could have a large impact in cancer care. Thus, the work here will serve as a promising prognostic tool for informing treatment choice and fostering therapy for lung cancer patients.

**Proposed Methodology**

## Proposed Methodology

Considering and keeping in mind the detail survey review reported21, the research methods for the proposed CAD system are devised with new efforts to utilize the effectiveness of radiomic feature model approach within curvelet domain as follows:

**A.Image Processing**

In practice, noise is the common problem in CT imaging under low dose protocol13. Lot of studies have shown that noise on CT images is found to be Gaussian13. Such degradations have a significant impact on the image quality and as a result, it affects the accuracy of CAD system even the experts with sufficient experience are not be able to draw correct and useful information from the images. Additionally, feature extraction, analysis, recognition and quantitative measurements become difficult and unreliable due to poor quality of images. Thus, the denoising and enhancement of the medical images become prime requirements for many practical applications.

From literature studies, it was clear that though several studies are based on noise filtering in the wavelet domain, they fail to give better results in edgy region because of generation of large wavelet coefficients even at fine scales and repeated at scale after scale, for the edges in the image. On the other hand, the anisotropic feature having the expertise particularly in preservation of edgy region in the denoising process will be considered to suppress noise and lung image enhancement14. To improve the speed, the iterations of the diffusion filters are kept small but the diffusion factor is optimized with the trade-off between noise reduction and feature preservation.

### B. ROI Extraction

Many of the earlier works, employs predetermined thresh- old value to separate the lung from the surrounding anatomy. Nevertheless, there are also some works that employ optimal thresholding for this task. For example, Otsu's thresholding method is based on the idea of finding a threshold value that minimizes the within-class variance of resulting foreground and background classes15. The method is robust, and it gives reasonable thresholding results in a vast variety of cases. However, this method has been shown to break down for a certain range of object-to-background pixel population ratios. With respect to this point, the current work refined Ostu's thresholding employing fuzzy membership function. Subsequently, the edges of lung are restored to prevent the exclusion of peripheral nodules by applying rolling ball technique described by Silva Sousa et al16.

### C. Pattern Recognition

Feature extraction is a significant step in the development of an automated characterization of lung nodules. Therefore, in this work, radiomic features are extracted using curvelets proposed by E. Candes and D. Donoho17. Curvelets are waveforms designed as an alternative to the widespread wavelet transforms. Unlike wavelets, has varying degree of localization in orientation with respect to the scale. This property makes curvelet transform capable of multi-scale representation in many directions and positions at each length scale. Also, curvelet takes less number of coefficients for handling discontinuities18. These features of curvelets motivate the usage of curvelet in this work with an objective to efficiently capture the finest curves within the textured medical image with respect to different scales and orientations. In the proposed system, radiomics features includes tumor intensity histogram (e.g., high or low contrast), tumor shape (e.g., round or spiculated), texture patterns (e.g., homogeneous or heterogeneous), as well as descriptors of tumor location and relations with the surrounding tissues (e.g., near the heart). Each of these features will be calculated for three levels of resolution and for each curvelet matrix based on each radial 'wedge' (16 angles).

To reduce the dimensionality of the feature space and further reduce the FP rates, SVM is applied in this work to analyze the feature variance and select the relevant features. The SVM finds hyperplane that separates various classes of nodules using Support Vectors19. These support vectors are transformed to a feature space with a nonlinear transformation. The non-linear transformations are represented with kernel function. The kernel function is used to map the input space to feature space. Here, the Morlet Wavelet kernel19 given below, is used as kernel of SVM Classifier

$$K(x) = \cos(1.75x) \cdot e^{\left(x^2/2\right)} \tag{1}$$

### D. Evaluation Metrics

The accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve22 will be employed to as validation metrics to evaluate the performance of the classifier. Here, the sensitivity (true positive rate, TPR) represents the number of true positives divided by the total number of positive cases. The specificity to measure the number of false positive rate. The accuracy (ACC) is the proportion of true results in the population. Finally, the ROC curve to visually illustrate the sensitivity and FP rate for different threshold values.

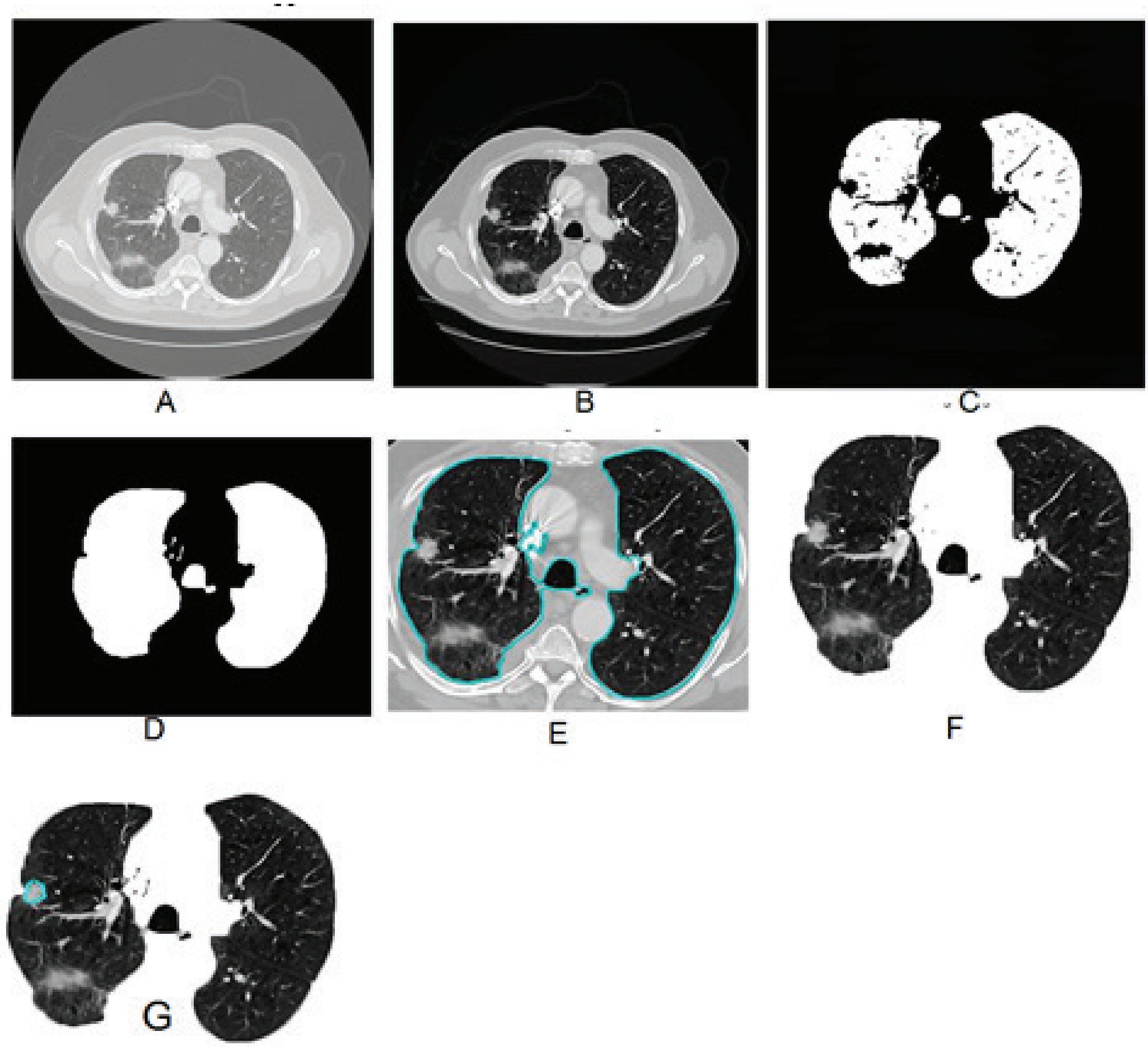


**Figure 1 Illustrates lung nodule detection process on slice number 45 of patient ID LIDC-IDRI-0191. (a) Original DCOM Image (b) Enhanced Image using diffusion filter domain (c) Segmented Binary Lung Image using Fuzzy K-means (d) Binary Image with lung boundary corrected using Rolling ball technique (e) boundary delineating the lung in Original Image (f) Segmented Lung Image (g) Lung Nodule**

## Experimental Design for Preliminary Study

A preliminary study was conducted to validate the feasibility of the proposed system adopting the publicly available LIDC database at in the National Biomedical Imaging Archive (NBIA)20. This database contains a greater number of subtle and irregularly shaped nodules. Also, the database contains ground truth information obtained from multiple radiologists. This dataset is composed of 1010 patients CT exams with 1500 nodules. LIDC case LIDC-IDRI- 0001 through LIDC case LIDC-IDRI-0150 was considered to train the proposed CAD system. LIDC case LIDC-IDRI- 0175 through LIDC case LIDC-IDRI-0200 was considered to test the proposed CAD system. During the experiments, the selected training images were also included in the testing stage for the proposed system. The ground truth information annotated by multiple radiologists was used to verify the correctness of the proposed system.

As, reduction of the feature vector is imperative to avoid SVM over-fitting and to obtain a manageable feature space, the preliminary experiments validated the system for several resolution levels and support vector kernels. Initially, three levels of resolution were investigated. A feature vector averaged over co-occurrence directions (99 total features) was compared with a feature vector averaged over both co- occurrence directions and curvelet details (33 descriptors). The feature vector averaged over both co-occurrence directions and curvelet details was discovered to be ideal. The following sets of descriptors were calculated: features on individual levels of resolution (1, 2, or 3, each with only 11 descriptors), features based on two levels.

## Results and Discussion

The results obtained from preliminary experimental set- tings are shown in Fig. 1 for the patient-id LIDC-IDRI-0191 and slice number 45. Although the preliminary results of our proposed system are encouraging and is expected to improve the specificity and sensitivity in Lung cancer diagnosis, we need more cases to test for generality. We will continue to perform additional testing and training with CT cases comprising nodules of various sizes and shapes available in the image database I-ELCAP. More importantly, extensive performance comparison between proposed system and radiologists will performed to evaluate the role of this system in clinical practice.

Next, since the nodules detected with the proposed CAD system were included in the reference standard, the study results could be potentially biased toward making the proposed CAD appear more accurate than it really is. But at this point, we are unable to confirm or reject the influence of this potential bias based on our preliminary study results. Therein, CT cases with complex nodule structures will be identified and considered for testing. The detection of such abnormal nodules may require a modification of current CAD implementation or an introduction of new algorithms or fine-tuning of parameters for the current system.

In future work, focus of attention will be put to improve the clinically relevant performance measures, sensitivity- specificity calculus of the proposed CAD systems. In this regard, systematic study will be performed to assess the relative weights and valid ranges of parameter values on the detection performance of the proposed system. We have planned to evaluate the ability of the classifier in the task of differentiating nodule candidates that correspond to actual nodules from false-positive candidates.

**Ethical Clearance:** The proposed work is a retrospective study that uses images from the public database archives. Therefore, there was no requirement to obtain any Informed consent from the institutional review boards.

**Source of Funding:** Self

**Conflict of Interest:** Nil

## References

[1]. Fintelmann FJ, Bernheim A, Digumarthy SR, Lennes IT, Kalra MK, Gilman MD, Sharma A, Flores EJ, Muse VV, Shepard JA. The 10 pillars of lung cancer screening: rationale and logistics of a lung cancer screening program. Radiographics. 2015 Nov;35(7):1893-908

[2]. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA: a cancer journal for clinicians. 2015;65(1):5-29

[3]. Nishio M. Computer-Aided Diagnosis of Lung Nodules: Systems for Estimation of Lung Cancer Probability and False-Positive Reduction of Lung Nodule Detection. in Lung Imaging and CADx 2019 Apr 24 (pp. 107-132). CRC Press.

[4]. Masood A, Yang P, Sheng B, Li H, Li P, Qin J, Lanfranchi V, Kim J, Feng DD. Cloud-Based Automated Clinical Decision Support System for Detection and Diagnosis of Lung Cancer in Chest CT. IEEE Journal of Translational Engineering in Health and Medicine. 2019 Dec 4;8:1-3

[5]. Robbins HA, Berg CD, Cheung LC, Chaturvedi AK, Katki HA. Identification of candidates for longer lung cancer screening intervals following a negative low-dose computed tomography result. JNCI: Journal of the National Cancer Institute. 2019 Sep 1;111(9):996-9

[6]. Al Mohammad B, Hillis SL, Reed W, Alakhras M, Brennan PC. Radiologist performance in the detection of lung cancer using CT. Clinical radiology. 2019 Jan 1;74(1):67-75

[7]. Chiesa-Estomba CM, Echaniz O, Larruscain E, Gonzalez-Garcia JA, Sistiaga-Suarez JA, Graña M. Radiomics and Texture Analysis in Laryngeal Cancer. Looking for New Frontiers in Precision Medicine through Imaging Analysis. Cancers. 2019 Oct;11(10):1409

[8]. Wakabayashi T, Ouhmich F, Gonzalez-Cabrera C, Felli E, Saviano A, Agnus V, Savadjiev P, Baumert

TF, Pessaux P, Marescaux J, Gallix B. Radiomics in hepatocellular carcinoma: a quantitative review. Hepatology international. 2019 Sep 1:1-4

[9]. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer. 2012 Mar 1;48(4):441-6.

[10]. Tian J, DI D, Wang S. Radiomics for lung cancer. Radiomics and Radiogenomics: Technical Basis and Clinical Applications. 2019 Jul 9:251

[11]. Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-scale convolutional neural networks for lung nodule classification. International Conference on Information Processing in Medical Imaging 2015 Jun 28 (pp. 588-599). Springer, Cham

[12]. Kumar D, Chung AG, Shaifee MJ, Khalvati F, Haider MA, Wong A. Discovery radiomics for pathologically-proven computed tomography lung cancer prediction. in International Conference Image Analysis and Recognition 2017 Jul 5 (pp. 54-62). Springer, Cham

[13]. Wolterink JM, Leiner T, Viergever MA, Išgum I. Generative adversarial networks for noise reduction in low-dose CT. IEEE transactions on medical imaging. 2017 May 26;36(12):2536-45

[14]. Fung YH, Chan YH. Optimizing the error diffusion filter for blue noise halftoning with multiscale error diffusion. IEEE transactions on image processing. 2012 Aug 2;22(1):413-7

[15]. Otsu N. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics. 1979 Jan;9(1):62-6

[16]. da Silva Sousa JR, Silva AC, de Paiva AC, Nunes RA. Methodology for automatic detection of lung nodules in computerized tomography images. Computer methods and programs in biomedicine. 2010 Apr 1;98(1):1-4

[17]. Starck JL, Candès EJ, Donoho DL. The curvelet transform for image denoising. IEEE Transactions on image processing. 2002 Aug 7;11(6):670-84

[18]. Starck JL, Candès EJ, Donoho DL. The curvelet transform for image denoising. IEEE Transactions on image processing. 2002 Aug 7;11(6):670-84

[19]. Maboudou-Tchao EM. Wavelet Kernels for Support Matrix Machines. Modern Statistical Methods for Spatial and Multivariate Data 2019 (pp. 75-93). Springer, Cham

[20]. Armato III SG, McLennan G, McNitt-Gray MF, Meyer CR, Yankelevitz D, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, Reeves AP. Lung image database consortium: developing a resource for the medical imaging research community. Radiology. 2004 Sep;232(3):739-48.

[21]. El-Regaily SA, Salem MA, Abdel Aziz MH, Roushdy MI. Survey of computer aided detection systems for lung cancer in computed tomography. Current Medical Imaging. 2018 Feb 1;14(1):3-18

[22]. Prabukumar M, Agilandeeswari L, Ganesan K. An intelligent lung cancer diagnosis system using cuckoo search optimization and support vector machine classifier. Journal of Ambient Intelligence and Humanized Computing. 2019 Jan 29;10(1):267-93