

## Lung Cancer Detection and Classification based on Image Processing and Statistical Learning

<sup>1</sup>Md Rashidul Hasan and <sup>2</sup>Muntasir Al Kabir

<sup>1</sup>Department of Mathematics and Statistics

The University of New Mexico Albuquerque, New Mexico, USA

<sup>2</sup>Anderson Department of Marketing, Information Systems, Information Assurance, and Operations Management,

The University of New Mexico Albuquerque, New Mexico, USA

**Corresponding Author: Md Rashidul Hasan.**

---

### Abstract

Lung cancer is one of the death threatening diseases among human beings. Early and accurate detection of lung cancer can increase the survival rate of a lung cancer patient. Computed Tomography (CT) images are commonly used for detecting the lung cancer. Using a data set of thousands of high-resolution lung scans collected from "Kaggle" competition, we have pre-processed the images by smoothing, enhancement, and segmentation. Then, we measured area, perimeter, skewness, kurtosis, and entropy from the CT images of lung. Finally, we have applied to several statistical methods such as logistic regression, Quadratic discriminant analysis, K-nearest neighbors, decision tree, random forest, k-mean clustering, and Support vector machine (SVM) to accurately determine the cancerous cell. These methods have been tested on 198 slices of CT images of various stages of cancer obtained from "Kaggle" data set and found satisfactory results. Among them SVM identified the cells (whether cancerous or not) with the highest accuracy of 72.2%. In terms of accuracy, random forest (71.2%) and decision tree (71.7%) are also very close to SVM for detecting the cancerous cells.

---

**Keywords:** Lung Cancer Detection, Image Processing, and Statistical Learning

---

### INTRODUCTION

Lung cancer is one kind of disease that grows uncontrolled way and form abnormal cells in the lung. These cells do not function like other normal cells. Because of DNA mutation by different factors like smoking, air pollution, Inherited gene changes, cancer can grow in human lungs. According to American Cancer Society, among all new cancers about 14% are lung cancers. They also estimate in 2018, there are about 234,030 new lung cancer in United States and about 154,050 deaths because of lung cancer [1]. Now a days, the reason of death is far beyond than prostate, colon, and breast cancers combined to lung cancer.

Objective of this study is to detect lung cancer using image processing techniques. CT scanned lung images of cancer patients are acquired from Kaggle Competition dataset [2]. Using image processing techniques like preprocessing, Segmentation and feature extraction, area of interest is separated. Developing the algorithm, features like area, perimeter and entropy are extracted from all the images. The parameter values obtained from these features are compared with the normal values suggested by a physician. From the comparison result, cancer nodules are detected. A graphical user interface is developed to scan all the images and display the features and cancer nodules. This system can help in early detection of lung cancer more

accurately. R. Shojaii [3] presented lung segmentation technique using watershed transform along with internal and external marker. They also used rolling ball filter for the smoothing of the contour and to fill the cavities of the cancer nodules. B. Hochegger [4] used genetic algorithm to select particular features and GLCM for the extraction. They used Support vector machines (SVM) to classify stages of lung cancer. A. kaur [5] investigated several existing literatures and identify the bilateral filter is best solution in using marker based watershed segmentation algorithm.

S. A. Amjed [6] proposed a computer aided detection (CAD) system to detect the lung cancer areas from CT images. S. Ignatious [7] used classifier such as SVM, NBM, NB Tree and Random Tree on 200 slices of CT images of several stages of cancer and found 94.4% accuracy while using Random Tree. I. Bush [8] presented ResNet CNN model to identify and localized different nodules. However, this model does not for localizing exact position nodule. F. Ciompi [9] applied a deep learning system to different dataset, one from Italian MILD screening trial as training data and another from the Danish DLCST screening trial as test data of lung cancer patients to compare the difference between computer and human as an observer. B. N. Narayanan [10] used three different training data based on sliced thickness in computer aided detection.

**MATERIALS AND METHODS**

In figure 1 step by step procedures for CT image analysis is shown which will be discussed in details in the following sections.



Figure 1 CT Image analysis steps for breast cancer detection

**A. Image Processing**

**1) Image Acquisition:** Because of low noise and better clarity, CT scan images of Lung cancerpatient are more useful compared to MRI and X-ray. Due to its lesser distortion property, CT scan is easier to handle for the preprocessing part. For our research work, the CT images has been acquired from Kaggle competition dataset. Now a days, DICOM (Digital Imaging and Communication in Medicine) is a standard format for medical imaging. Figure2(a) shows a typical CT image of lung cancer patient used for analysis. The acquired images are in the raw form and observed a lot of noise.

**2) Pre-Processing**

**a) Smoothing:** To improve the contrast, clarity, separate the background noise, it is required to pre-process the images. Hence, various techniques like smoothing, enhancement are applied to get image in required form. It suppresses the noise or other small fluctuations in the image; equivalent to the suppressions of high frequencies in the frequency domain. Smoothing also blurs all sharp edges that bear important information about the image. To

remove the noise from the images, median filtering is used. Median filtering is a non-linear operation often used in image processing to reduce salt and pepper noise. In general, the median filter allows a great deal of high spatial frequency detail to pass while remaining very effective at removing noise on images where less than half of the pixels in a smoothing neighborhood have been affected.  $B = \text{medfilt2}(A, [m, n])$  performs median filtering of the matrix  $A$  in two dimensions. Each output pixel contains the median value in them  $x \times n$  neighborhood around the corresponding pixel in the image.

**b) Enhancement:** Enhancement technique is used to improve the interpretability or perception of information in images for human viewers, or to provide better input for other automated image processing techniques. Image enhancement can be classified in two main categories, spatial domain and frequency domain. Here histogram equalization is used for enhancement purpose and the output after performing enhancement from original image is shown in figure 2(b).

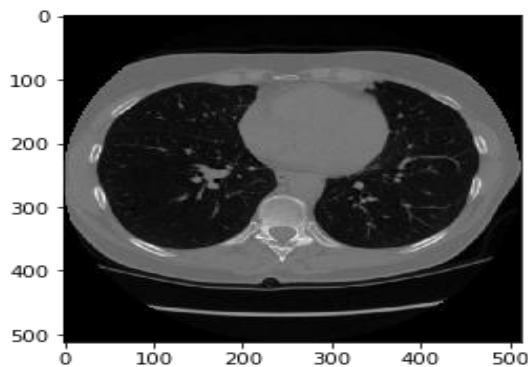
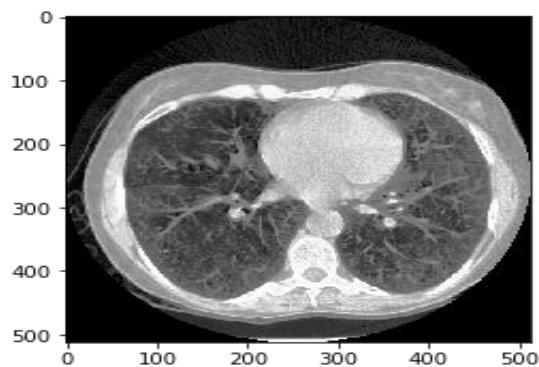


Figure 2 a) original image



b) enhanced image

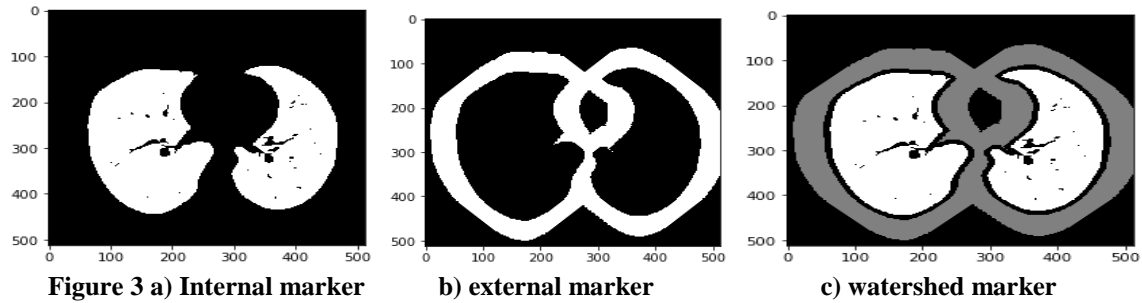
**B. Segmentation**

Image segmentation is a process of subdividing an image into the constituent parts or objects in

the image. So the main purpose of subdividing an image into its constituent parts or objects present in the image is that we can further

analyze each of the constituents or each of the objects present in the image once they are identified or we have subdivided them. So resulted output of image segmentation is a

collection of segments of entire image. In our method we use marker-controlled watershed segmentation.



In order to use marker based watershed (figure 3(c)) segmentation, we use internal marker shown in figure 3(a), that is definitely lung tissue and an external marker shown in figure 3(b). to find the precise

border of the lung we also used the Sobel-GradientImage shown in figure 4 (a) of our original scan.

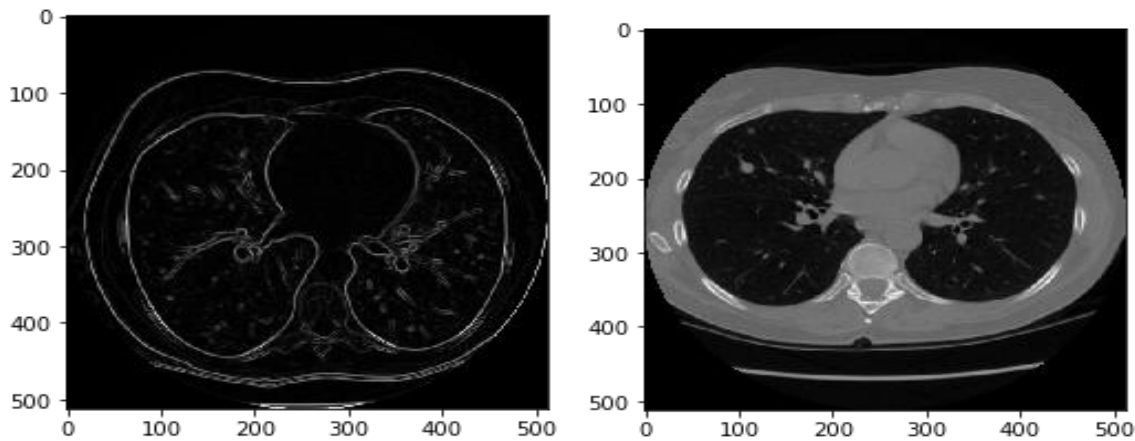


Figure 4 a) Sobel Gradient Image

**b) Segmented Lung Image**

The different steps involved in Marker Controlled Segmentation are the following:

- Step 1: Read in the color image and convert it to gray scale image.
- Step 2: Compute the Gradient Magnitude as the segmentation function.
- Step 3: Mark the foreground objects within the image.
- Step 4: Find out the background marker points within the image.
- Step 5: Find out the watershed transform of the segmented function of the image.
- Step 6: Resultant segmented binary image shown in figure 4(b) is obtained.

**C. Feature Extraction**

This stage is an important stage that uses algorithms and techniques to detect and isolate various desired portions or shapes segmented image. After labeling the segmented image we extracted the various features. The basic characters of feature are area, perimeter and eccentricity. These are measured in scalar. These features are defined as follows: **Area** is

one of the key parameters required for classification process. **Area** actually tells us about the size of the lump. **Perimeter**, another important parameter gives us the idea about the boundary of the defected cell. **Standard Deviation**,  $\sigma$  is the estimate of the mean square deviation of the grey scale pixel value from its mean,  $\mu$ . **Skewness** characterizes the degree of asymmetry of a pixel distribution in the specified ROI around its mean. **Kurtosis** measures the peakness or flatness of a distribution relative to a normal distribution. **Entropy** is a measure of the maximal amount of potential information given by the segmented ROI.

**CLASSIFICATION**

Before discussing the classification, we divide our data set into training and test data. The training data set consists of 1397 patients where 1035 patients do not have cancer and rest of 362 do have. On the other hand, our test data set contains 198 patients where 57 patients are carrying cancerous region and 141 without that region. In this section, we want to choose a model based on our training data and then test the model for accuracy. For choosing the model

we tried both supervised and unsupervised learning. However, our goal is to predict the response variable cancer (yes or no) which is a categorical variable. Moreover, we want to make sure that there is no problem of collinearity among the predictor variables. Figure 5 shows the scatter plot matrix of all

variables to check the collinearity. This matrix plot indicates high correlation in between the predictors area, entropy, and standard deviation. Furthermore, skewness and kurtosis are also highly correlated to each other. Therefore, this collinearity suggests us to eliminate some of our predictor variables.

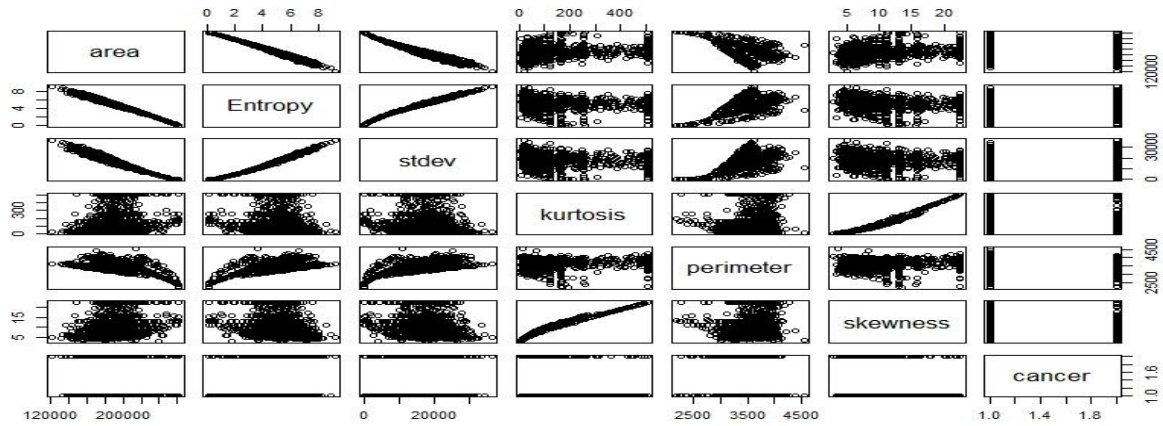


Figure 5 Scatter plot matrix of training data set

We used best subset selection method for eliminating non significant predictors. Then we applied different supervised and unsupervised learnings (James, 2014). The methods and classifications are discussed below.

**A. Best Subset selection**

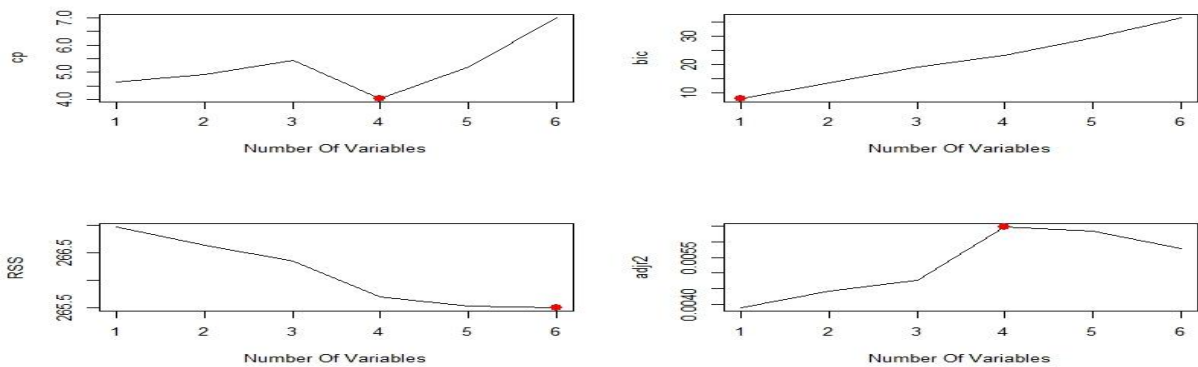


Figure 6 Plot for  $C_p$ , BIC, RSS, and adjusted  $R^2$

Then we tried four as well as three predictors separately and found that entropy, standard deviation and perimeter are statistically significant. However, for classification we tried two cases (i) all predictors and (ii) three predictors to see if there were any improvisation in accuracy level. We apply multiple logistic regression in the next section.

**B. Multiple Logistic Regression**

For various predictors  $X_1, X_2, \dots, X_p$  the multiple logistic regression is generalized as follows:

$$P(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

We ran a linear regression model for each possible combination of the X's. Fitting all models with  $k$  predictors where  $k = 0, 1, \dots, p$ , we selected the best model,  $M_k$ , using minimum RSS,  $C_p$ , BIC, or highest adjusted  $R^2$ . For our data set both  $C_p$  and adjusted  $R^2$  suggested 4 predictors as shown in figure 6.

Using all the predictors, this logistic regression method gave us no significant predictor variables except the standard deviation. However, this method predicted 60.1% data accurately. With the three predictors logistic regression model then gave us a improved accuracy level of 69.19%. Next section we apply linear discriminant analysis.

**C. Linear Discriminant Analysis (LDA) for  $p > 1$**

Assume that  $X = X_1, X_2, \dots, X_p$  is drawn from a multivariate normal distribution, with a class-specific multivariate mean vector and a common covariance matrix. Therefore,

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T (\Sigma)^{-1} (x - \mu)\right)$$

Then the Bayes classifier assigns an observation  $X = x$  to the class for which

$$\delta_k(x) = x^T (\Sigma)^{-1} \mu_k - \frac{1}{2} \mu_k^T (\Sigma)^{-1} \mu_k + \log(\pi_k)$$

Unfortunately, this method did not work. One of the reasons might be the relationship between the response and predictors are not linear. For example, figure 7 shows the curvilinear relation between cancer and entropy. Next, we apply quadratic discriminant analysis.

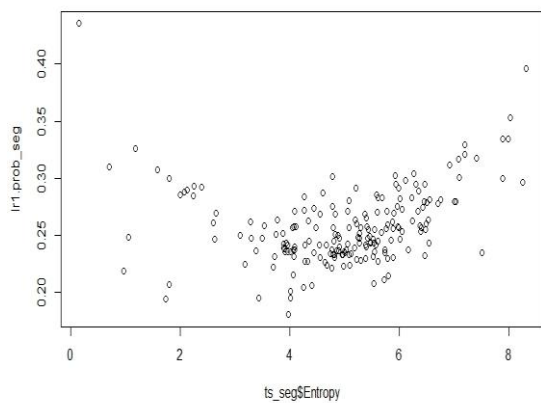


Figure 7 Relation between cancer and entropy

**D. Quadratic Discriminant Analysis (QDA)**

Again, assume that  $X = X_1, X_2, \dots, X_p$  is drawn from a multivariate normal distribution. Therefore, the Bayes classifier assigns an observation  $X = x$  to the class for which

$$\delta_k(x) = -\frac{1}{2} x^T (\Sigma)^{-1} x + x^T (\Sigma)^{-1} \mu_k - \frac{1}{2} \mu_k^T (\Sigma)^{-1} \mu_k - \frac{1}{2} \log(|\Sigma|) + \log(\pi_k)$$

For QDA, all predictor variables gave us 69.69% accuracy and when used three predictors we got slightly higher accuracy level of 71.21%. Next, we apply K-nearest neighbors Regression.

**E. K-nearest neighbors (KNN) Regression**

To predict Y for a given X value, consider the K closest points to X in training data and take the average of their response,

$$f(X) = \frac{1}{k} \sum_{X_i \in N_i} Y_i$$

Where  $N_i$  is the set of K neighbors based on X. For KNN, all predictor variables gave us 62.12% accuracy and when used three predictors we got slightly higher accuracy level of 64.64%. Next, we apply classification trees.

**F. Classification Trees**

A classification tree is used to predict a qualitative response rather than a quantitative one. For a classification tree, we predict that each observation belongs to the most commonly occurring class of

training observations in the region to which it belongs. In interpreting the results of a classification tree, we are often interested not only in the class prediction corresponding to a particular terminal node region, but also in the class proportions among the training observations that fall into that region. A large tree with lots of leaves tends to overfit the training data. We may consider reducing the tree by “pruning” some of the leaves. This can lead to improved accuracy. We use cross-validation to check which tree has the lowest RSS or error rate. Fortunately, there is software in place to perform all these calculations.

Because of some computational complexity we could not use all the training data for classification trees. However, we managed to handle 600 observations. Then, all predictor variables gave us 71.71% accuracy with 8 nodes as shown in figure 8 and after pruning, 3 nodes had been used but we got exactly same accuracy level.

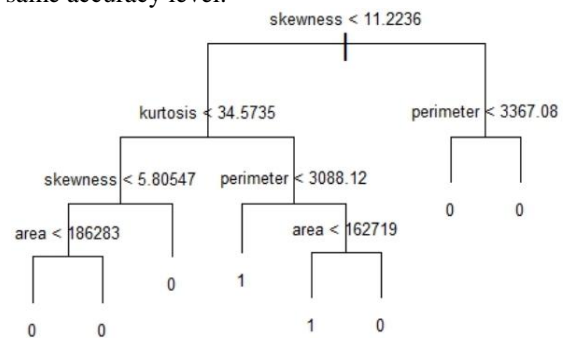


Figure 8 Classification tree using 8 nodes  
Next, we apply random forest method.

**G. Random Forests**

Random forest is a very efficient statistical learning method. It builds on bagging (in bagging, we build a number forest of decision trees on bootstrapped training samples) but provides an improvement because it de-correlates the trees. Build a number of decision trees on bootstrapped training samples. Each time a split in a tree is considered, a random sample of  $m$  predictors is chosen as split candidates from the full set of  $p$  predictors, typically  $m = \sqrt{p}$ . Suppose that there is a very strong predictor. The rest are moderate to poor predictors. For the bagged trees, most of the them will have the strong predictor for the first split. All bagged trees will look similar and the respective predictions, highly correlated. Averaging highly correlated quantities does not help with variance reduction. Random forests de-correlate the bagged trees.

Figure 9 represents the classification when used the random forest. Surprisingly the accuracy level is as same as the classification trees i.e. 71.71%.

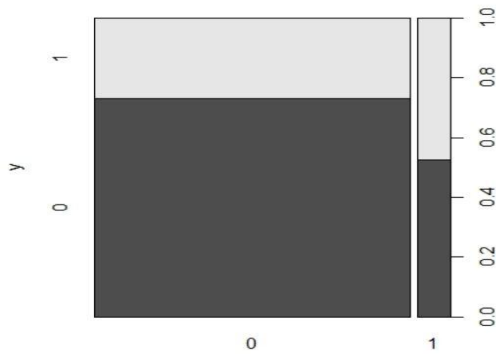


Figure 9 Classification using random forest

**H. Support Vector Machine (SVM)**

The support vector classifier finds the optimal hyperplane in the space spanned by  $X_1, X_2, \dots, X_p$ . In support vector machine we create a transformation  $b_1(X), b_2(X), \dots, b_m(X)$ . Then we find the support vector classifier in the transformed space. This produces a linear support classifier in  $b_1(X), b_2(X), \dots, b_m(X)$  but its non linear in the original  $X_1, X_2, \dots, X_p$ . In general, a support vector machine can be expressed through kernels as

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(X, X_i)$$

where  $K(.,.)$  is a Kernel function between two vectors. We define two vectors as  $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$  and  $x_l = (x_{l,1}, x_{l,2}, \dots, x_{l,p})$  then the Possible kernels are (i) inner product kernel is  $K(X_i, X_l) = \sum_{j=1}^p X_{i,j} X_{l,j} = \langle X_i, X_l \rangle$  (ii) polynomial kernel is  $K(X_i, X_l) = (1 + \sum_{j=1}^p X_{i,j} X_{l,j})^d$ , and (iii) radial kernel ( $\gamma > 0$ ) is  $K(X_i, X_l) = \exp(-\gamma \sum_{j=1}^p X_{i,j} X_{l,j})^2$ .

Applying the dataset to SVM, we get 71.71% accuracy before tuning the cost and gamma parameters. Then we tune these two parameters and got the best results for cost=1 and gamma=1. This moderately improved our accuracy level to 72.22%. Two predictors, area and perimeter have been used for SVM as shown in figure 10.

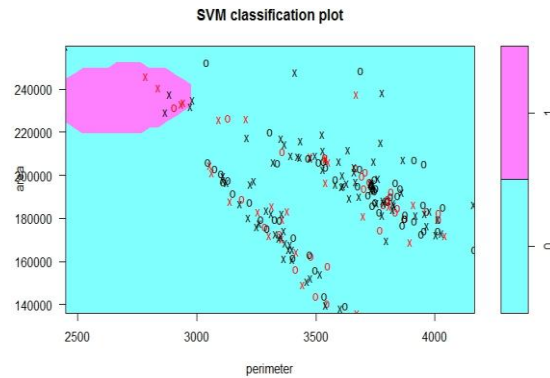


Figure 10 Classification using support vector machine

Finally, K-means clustering also applied in the next section.

**I. K-Means Clustering**

K-means clustering is a simple and elegant approach for partitioning a data set into K distinct, non-overlapping clusters. To perform K-means clustering, we must fix the desired number of clusters K. Suppose  $C_1, C_2, \dots, C_K$  are indices of the observations that define each cluster where  $C_k \cap C_{k'} = \emptyset$ , i.e. sets are mutually exclusive. Let  $W(C_k)$  measures how much observations differ within a cluster. The goal is to select  $C_1, C_2, \dots, C_K$  so that they minimize

$$\sum_{i=1}^k W(C_i)$$

Where  $W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \|X_i - X_{i'}\|^2$ , here  $x_i$

is the vector of all covariates for observation  $i$ ,  $|C_k|$  is the total number of elements in  $C_k$ . In this formulation,  $W(C_k)$  depends on the mean of each variable  $X_j$  for  $C_k$  (centroids). When used all predictors k-means clustering for training data gave 52.97% accuracy and for three predictors we got 54.67%. For test data using all predictors gave the accuracy level of 47.47% and three predictors gave slightly improved level of 55.05%. Figure 11 shows the k-means clustering for area and perimeter.

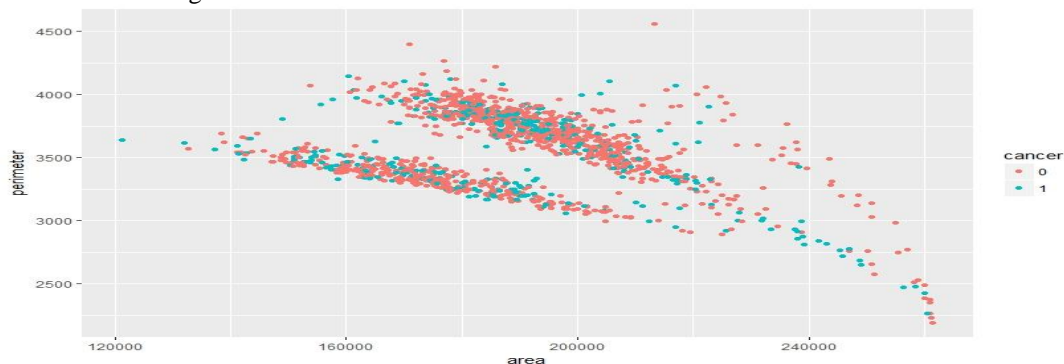


Figure 11 Classification using k-means clustering

## RESULTS

Figure 12 represents the summary of accuracy level. Blue and orange color indicates the the percentage of accuracy for all predictors and three predictors respectively. In contrast, different colors for SVM is

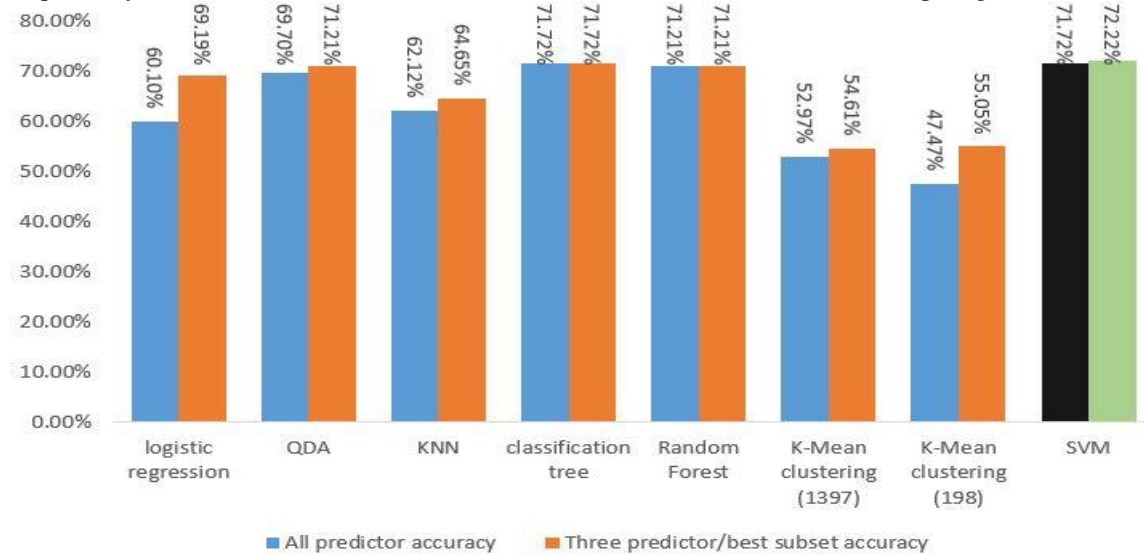


Figure 12 Bar graphs comparing the accuracy level

## Future Enhancement

In the proposed system we used only watershed marker-based segmentation in image processing part (I. Levner, 2007). In Future work, we want to use some other segmentation technique and compare. We believe that will increase our extracted feature quality. We also considering using some other filter and image enhancement method. The accuracy can be increased by extracting more features of the tumor, increasing the size of the dataset.

## CONCLUSION

The proposed lung cancer detection identifies the tumor within the lung. The CT image is pre-processed, and the pre-processed image is then subjected to segmentation by using Marker Controlled watershed segmentation. Segmented image is used for feature extraction. With the extracted features the tumor is detected within the lung. Both supervised and unsupervised classifier is used for the identifying of the cancer. The accuracy rate of the proposed system is 72.2% by using support vector machine. Thus, this system helps the radiologist to identify the stage of the tumor and increase the accuracy.

## REFERENCES

[1]. Robert A. S., Kimberly S. A., Brooks D., Stacey A. F., Manassaram-Baptiste D., Saslow D., Otis W. B. and Richard C. W., "Cancer screening in the United States, 2018: A review of current American Cancer Society guidelines and current issues in cancer screening," *CA: A Cancer Journal for Clinicians*, pp. 297-316, (2018).

for two different cost and gamma parameters. This bar graph also shows that SVM provides us the highest accuracy level while QDA, classification tree, and random forest are competing with SVM.

[2]. Kaggle, "https://www.kaggle.com/c/data-science-bowl-2017," Kaggle, (2017).

[3]. Shojaii R., Alirezaie J. and Babyn P., "Automatic lung segmentation in CT images using watershed transform," *IEEE International Conference on Image Processing 2005*, pp. II-1270, (2005).

[4]. Hochhegger B., Alves G. R., Irion K. L., Fritscher C. C., Fritscher L. G., Concatto N. H. and Marchiori E., "PET/CT imaging in lung cancer: indications and findings," *Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisiologia*, p. 264-274, (2005).

[5]. kaur A., Verma A. and Ssiet D., "The Marker-Based Watershed Segmentation- A Review," *International Journal of Engineering and Innovative Technology*, (2013).

[6]. Amjed S. A., Eslam B. J. and Mohammed A. A., "Automated detection of lung cancer using statistical and morphological image processing techniques," *Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisiologia*, (2014).

[7]. Ignatious S., Joseph R., John J. and Prahladan A., "Computer Aided Lung Cancer Detection and Tumor Staging in CT image using Image Processing," *International Journal of Computer Applications*, pp. 29-33, (2015).

[8]. Bush I., "Lung Nodule Detection and Classification," (2016).

- [9]. Ciompi F., Chung K., Van Riel S. J., Seito A. A. A., Gerke P. K., Jacobs C., Scholten E. T., Schaefer-Prokop C., Wille M. M. W. and Marchianò A., "Towards automatic pulmonary nodule management in lung cancer screening with deep learning," *Scientific Reports*, (2017).
- [10]. Narayanan B. N., Hardie R. C. and Kebede T. M., "Performance analysis of a computer-aided detection system for lung nodules in CT at different slice thicknesses," *Journal of medical imaging* (Bellingham, Wash.), p. 014504, (2018).
- [11]. James G., Witten D., Hastie T. and Tibshirani R., *An Introduction to Statistical Learning: With Applications in R*, Springer Publishing Company, Incorporated, (2014).
- [12]. Levner I. and Zhang H., "Classification-Driven Watershed Segmentation," *IEEE Transactions on Image Processing*, pp. 1437-1445, (2007).