An Enhanced Approach for Detection and Classification of Computed Tomography Lung Cancer

Wafaa Alakwaa
Faculty of Computers & Info.
Cairo University, Egypt

Mohammad Nassef
Faculty of Computers & Info.
Cairo University, Egypt

Amr Badr
Faculty of Computers & Info.
Cairo University, Egypt

***Abstract*—The paper presents approaches for nodule detection and extraction in axial lung computed tomography. The goal is to detect correctly pulmonary nodule to recognize and screen lung cancer patients. The pulmonary nodule detection is very challenging problem. The proposed model developed a hybrid efficient model based on affine-invariant representation and shape of segmented nodule. Due to large number of extracted features for all slices on patient, feature selection is an important step to select the most important feature for classification. We apply forward stepwise least squares regression that maximizes the R-squared value, this criterion provides a fast preprocessing feature selection assessment for systems with huge volumes of features based on a linear models framework. Moreover, gradient boosting have been suggested to select the relevant features based on boosting approach. Classification of patients has been done by support vector machine. Kaggle DSB dataset is used to test the accuracy of our model. The results show major improvement in accuracy and the features are reduced.**

Keywords*—*Lung cancer; computed tomography; affine invariant moments; pulmonary nodules; R2; feature selection; support vector machine

I. Introduction

Lung cancer has the second highest incidence of cancers worldwide for both the male and female population, and remains the cancer with the highest mortality. This is because it remains asymptomatic for a long time, and is therefore diagnosed mostly at such a late stage that treatment outcome is poor. Despite this, most countries currently do not have a lung cancer screening programme for early detection of lung cancer. This is not only due to the high costs involved if applied to a large proportion of the population, but also the lack of a sufficiently sensitive diagnostic test, including imaging. Current research in screening for lung cancer is therefore limited to patients identified at high risk of developing lung cancer, such as smokers or patients with COPD (or both), but it is anticipated that this research could form an important foundation for a future national screening programme [1].

CT scan is an extended version of X-ray in which computer is attached to the X-ray machine. Pictures that are taken from angles and distances are processed in the computer and presented in the 3-dimensional, cross-sectional (tomographic) and in slices form. In this way, bones, tissues, blood vessels, and organs are shown up clearly. The imaging of CT scan is useful for diagnosis, treatment and progress of medication. Recently, helical or multi-slice scanning is introduced that almost eliminated gaps in the collection of slides [2]. The radiologists miss detecting lung nodules in early stage due to dramatic expanding in number of image slices in high resolution images. A lung Cancer screening computer-aided detection/diagnosis (CAD) system can reduce cost and speed up screening. CAD systems help radiologists in building decisions and enhance process of detection and observation of diseases in screening. CAD can enhance nodule detection step by detecting missed nodules, reduce reading time so that the screening process is made possible and helps differentiate between benign and malignant lesions.

In this paper, we introduce an efficient model to detect and diagnose lung cancer patients. Based on watershed segmentation, nodules are detected and shape features are applied to describe the nodules using affine moments. Gradient boosting is used which can identify a robust feature selection through ensemble learning by combining weak classifiers to yield strong, robust and accurate classifier. The variations in the target classes are identified by the best selected features through R-Squared regression criterion.

The paper's arrangement is describes as follows: Related work is summarized briefly in Section II. The model architecture is presented in Section III. The nodule segmentation is introduced in Section IV based on watershed algorithm. The feature extraction process based on affine moments and shape features are presented in Section V. Section VI presents feature selection models based on ensemble-based feature selection models which include Gradient Boosting and regression-based feature selection using R-squared model. The classification process using SVM is mentioned in Section VII. Our discussion and results are described in details in Section VIII. Section IX summarizes the conclusion of paper.

II. Related Work

Recent research tries to encourage developing an image-based model that is able to improve, as a second opinion, in conjunction with the radiologist, the detection accuracy of a radiologist,and reduce mistakes related to false positives. A CAD system generally consists of several steps when processing medical images. Images are preprocessed to remove

noise and enhance quality. Then Region of Interest is segmented from other structures. Features are extracted from these ROIs ,such as geometrical, textural, and statistical features. Accordingly, a classification step is done, to decide if the image contains malignant nodule. There has been exhastive efforts on computer aided diagnosis for lung images.

In [3], [4], Hamada et al. evaluated their system on the Japanese Society of Radiological Technology (JSRT) standard dataset of chest radiographs. The two preprocessing techniques were histogram equalization and Laplacian filter. Contrast was enhanced and the rapid intensity change was examined. Wavelet transform was used for feature extraction. To select the most important features the proposed model calculated the variance and the energy. The dimensions of the overall features is then reduced. For classification K-nearest neighbor classifier was employed. The proposed model was tested on 154 nodule regions with 100 malignant and 54 benign nodules. The Accuracy was 99.15% for normal versus abnormal and 98.70% for benign and versus malignant.

In [5] many techniques were applied for lung region detection. Bit plane slicing algorithm is used to generate different binary slices which then were enhanced by erosion algorithm and dilation and median filters. After detection of lung region, segmentation was applied to identify the lung nodules. Fuzzy Possibilistic C Mean (FPCM), which is a clustering algorithm that combines the characteristics of a fuzzy and possibility c-means, was applied for segmentation. Area and the mean intensity value of the candidate region are the features that were used to classify the nodule on. Support Vector Machine was used for binary classification. The proposed model was tested on experimentation data consists of 1000 lung images obtained from the reputed hospital.

In [6], Ada and Rajneet K. proposed a hybrid approach on feature extraction and Principal Component Analysis (PCA). Histogram Equalization is used for preprocessing of the images. Features were Extracted using Binarization and Masking Approach. A Grey Level Co-occurrence Method was created to make different combinations of pixel brightness. The features used in this approach were entropy, contrast, energy, and maximum probability. The exact output and results were not clearly specified.

In [7], FFT, Auto enhancement and Gabor filtering were used for image enhancement step. Topology surface and watershed algorithm were applied to the marker location and segmentation progress. The features that were extracted from ROI were area, perimeter, eccentricity and average intensity. In [8] Kamil Dimililer et al. used many image processing techniques: grayscale conversion, thresholding, erosion, median filtering and image subtraction.

In [9], many filters were applied in the preprocessing step, such as low pass filters, contrast stretching histogram equalization, negativity and power law transformation. For segmentation modified thresholding, labeling algorithm and edge detection were taken off. Features such as geometric properties, textural properties and mathematical properties were calculated. Gray Level Co-occurrence Matrix (GLCM) is a used to examine relationship of image pixels.

In [10] a computer aided diagnosing system was proposed to detect lung cancer based on texture features take out from the slice of DICOM Lung CT images. For preprocessing step K Nearest Neighbors and Weiner filters were used. Sobel Methods was suggested for segmentation. The set of texture features that were used for diagnosis are area of the interest, Calcification, Shape, Size of nodule and Contrast Enhancement. Artificial neural network was used for classification. This CAD system neglects all the false positive cancer regions and detects the cancer regions. The used dataset was obtained from NIH/NCI Lung Image Database Consortium (LIDC). There were about 1000 lung images. This approach showed sensitivity of 90% with 0.05 false positives per image.

In [11], Hashemi et al. proposed a system based on fuzzy inference. Starting with image enhancement and noise removal, Linear-Filtering was used. A region growing based technique was used for segmentation. A Fuzzy Inference System was implemented to determine the type of the mass diagnosed. The system was 95% accurate. Features such as area and color were used. This method was tested on 1000-tumor contained 10000 CT slices from 1000 lung tumor patients. The accuracy of the proposed systems was 95%.

III. Proposed Methodology

The suggested framework of Lung cancer detection and classification is composed of four stages: Nodule segmentation, feature extraction, feature selection and patient classification. As shown in Fig. 1, the overall architecture is drawn. Watershed segmentation is used to detect the nodule in lung cancer slices in CT scan as feature detection. Then, shape features and invariant affine moments are applied to describe the extracted nodules. For feature selection, we developed ensemble models and regression model to select the best important and relevant features to avoid the over-fitting problems. Finally, patient is classified by SVM. The main tasks of our model are presented in details through the next sections.

IV. Nodule Detection and Segmentation

One of the main tasks in medical diagnosis is the segmentation, especially in lung cancer using CT scans. Segmentation is a commonly preprocessing step for more enhancement in anomalies and lung structures, such as nodules.

The watershed algorithm is a common segmentation technique based on morphology mathematics. It depends on an intensity based topographical representation. The higher altitudes (hills) are represented by brighter pixels and the valleys are represented by dark pixels to determine the path of a falling raindrop would follow. The different regions are separated by watershed lines in watershed algorithm. Fig. 2 shows the resulted nodules extracted using watershed.

V. Feature Extraction

A. Moments Invariant Features

Moments are applied in many applications. Many of these techniques are essentially based on the general moment theory widely known and applied in research in several areas of statistics and mechanics. In particular, geometric moments have vast practical applications in many area of computer vision and invariant pattern recognition, ranging from lower-level recognition such as pose estimation to higher-level recognition such as

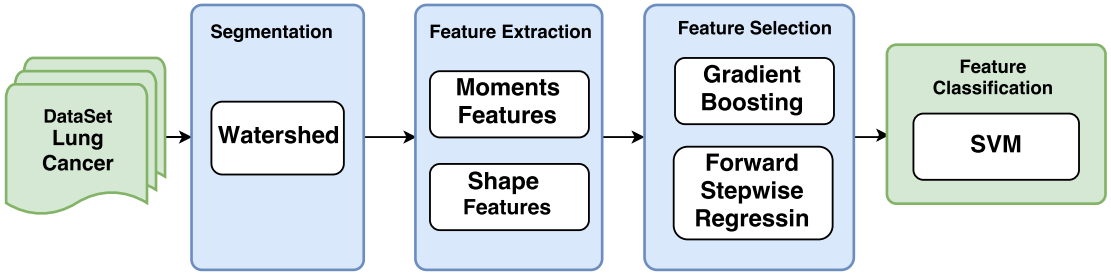

Figure 1: Pipline of the proposed model

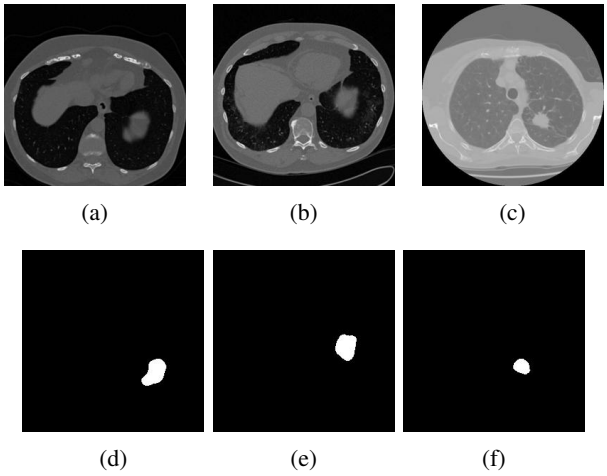

Figure 2: An axial slices of CT scans for Three Different Patients 2a, 2b, 2c and Segmented Nodules Based on Watershed Algorithm for Each Patient 2d, 2e, and 2f

activity recognition and analysis. When applied to images, they were identified to be most descriptive of the image contents (i.e., intensity distribution) with respect to its axes. Once such moments are properly defined, both global and detailed geometric information of image contents can be reasonably expected to be detected robustly. In such a scenario, moments would be able to characterize various image objects such that the properties with analogies in statistics or mechanics are extracted, and thus the shape of all objects of interest can be described well. Formally speaking, in continuous domain, an image is viewed as a 2-D Cartesian density distribution function f(x, y). The general form of the geometric moments of order (p + q) for the function f(x, y), evaluated over the entire plane Ω is defined by the following discrete form:

$$M_{pq} = \sum_{y} \sum_{x} \varphi_{pq}(x, y) I(x, y), \quad p, q = 0, 1, 2, ..., \infty \quad (1)$$

Where, φ_{pq} is a basis function or weighting kernel by which a weighted description for the image function f(x, y) across the entire plane Ω is generated. It is perhaps worthwhile to point out here that the choice of above basis functions φ_{pq} greatly depends on the application of use, and on the invariant properties desired. Furthermore, it is expected that choosing a specific basis function results in some constraints, such as to restrict the range of the image coordinates, x and y, enable the image and its descriptors to be translated to other coordinates (e.g., polar coordinates), etc. In [12], Hu stated that the 2-D Cartesian moment of order (p + q) for an $m \times n$ discretized image, I(x, y) can be defined by taking the basis function in (1) as a monomial of power p + q (product of powers of the variables x and y, i.e., $\varphi_{pq}(x, y) = x^p y^q$ as follows:

$$M_{pq} = \sum_{y=0}^{n-1} \sum_{x=0}^{m-1} x^p y^q I(x, y), \quad p, q = 0, 1, 2, ..., \infty \quad (2)$$

The full moment set of order k that includes all moments, M_{pq}, such that $p+q \geq k$ compromises of exactly $\frac{1}{2}(k+1)(k+2)$ elements. Ever since the pioneering work of Hu [12] on moment functions that has explored quite thoroughly the use of moments for image analysis and object representation, a broad range of new applications utilizing moment invariants in image analysis and pattern recognition fields has started to evolve. It is clear that the Cartesian moments given by (2) are not invariant to geometric transformations. To achieve invariance under translation, these moments are calculated with respect the center of mass as follows:

$$\mu_{pq} = \sum_{y=0}^{n-1} \sum_{x=0}^{m-1} (x - \bar{x})^p (y - \bar{y})^q I(x, y), \quad p, q = 0, 1, 2, ..., \infty \quad (3)$$

Where, $\overline{x}$ and $\overline{y}$ are the coordinates of the centroid and given by:

$$\bar{x} = \frac{M_{10}}{M_{00}}, \qquad \bar{y} = \frac{M_{01}}{M_{00}} \quad (4)$$

After a bit tedious but straightforward manipulation, (2) and (3) lead to the following relation between the Cartesian and centralized moments:

$$\mu_{pq} = \sum_{i}^{p} \sum_{j}^{q} \binom{p}{i} \binom{q}{j} (-\bar{x})^{p-i} (-\bar{y})^{q-j} M_{ij} \quad (5)$$

However, it should be emphasized that the expression in (3) suggests that the centralized moments are only invariant to translation. To enable invariance under scale changes, the

2-D centralized moments μ_{pq} need to be normalized to obtain scale-normalized centralized moments η_{pq} as follows:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{6}$$

Where, the exponent γ is given in terms of p and q as follows:

$$\gamma = \frac{p+q}{2} + 1, \quad p+q \geq 2$$

Strictly speaking, the moments present the shape properties for appearance of a nodule. Affine moments are invariant under six transform and derived based on central moments [13] as follows:

$$\begin{aligned}
I_1 &= \frac{1}{\eta_{00}^4}[\eta_{20}\eta_{02} - \eta_{11}^2], \\
I_2 &= \frac{1}{\eta_{00}^{10}}[\eta_{03}^2\eta_{30}^2 - 6\eta_{30}\eta_{21}\eta_{12}\eta_{03} + 4\eta_{30}\eta_{12}^3 \\
&\quad + 4\eta_{03}\eta_{21}^3 - 3\eta_{21}^2\eta_{12}^2], \\
I_3 &= \frac{1}{\eta_{00}^7}[\eta_{20}(\eta_{21}\eta_{03}\eta_{21} - \eta_{12}^2) - \eta_{11}(\eta_{30}\eta_{03} - \eta_{21}\eta_{12}) \\
&\quad + \eta_{02}(\eta_{03}\eta_{12} - \eta_{21}^2)], \\
I_4 &= \frac{1}{\eta_{00}^{11}}[\eta_{20}^3\eta_{03}^2 - 6\eta_{20}^2\eta_{11}\eta_{12}\eta_{03} - 6\eta_{20}^2\eta_{02}\eta_{21}\eta_{03} \\
&\quad + 9\eta_{20}^2\eta_{02}\eta_{12}^2 + 12\eta_{20}\eta_{11}^2\eta_{21}\eta_{03} + 6\eta_{20}\eta_{11}\eta_{02}\eta_{30}\eta_{03} \\
&\quad + 18\eta_{20}\eta_{11}\eta_{02}\eta_{30}\eta_{12} - 8\eta_{11}^3\eta_{30}\eta_{03} - 6\eta_{20}\eta_{02}^2\eta_{30}\eta_{12} \\
&\quad + 9\eta_{20}\eta_{02}^2\eta_{21}^2 + 12\eta_{11}^2\eta_{02}\eta_{30}\eta_{12} + \eta_{02}^3\eta_{30}^3], \\
I_5 &= \frac{1}{\eta_{00}^6}[\eta_{40}\eta_{04} - 4\eta_{31}\eta_{13} + 3\eta_{22}^2], \\
I_6 &= \frac{1}{\eta_{00}^9}[\eta_{40}\eta_{04}\eta_{22} - 4\eta_{31}\eta_{13}\eta_{22} - \eta_{40}\eta_{13}^2 - \eta_{04}\eta_{13}^2 - \eta_{22}^2]
\end{aligned} \tag{7}$$

B. Shape Features

After segmentation, the nodule candidate is selected and two different types of features are extracted, namely, 2-D geometric, 3-D geometric. A median slice $I_{NC,m}$ is extracted from 2-D features because the area of the segmented object is the largest. The shape of nodule candidates are worthy features to recognize the objects in Lung. The shape of nodules are described as 2-D and 3-D geometric features. Area, Perimeter, and Eccentricity are the most common used in our paper to describe the segmented regions in lung cancer slices.

VI. Feature Selection

Feature selection is a worthy stage in medical diagnosis to choose the best features that enhance the model accuracy. Furthermore, the models can be simpler and faster in understanding and building with the least number of features. In our paper, we applied two approaches: regression-based feature selection and tree-based feature selection.

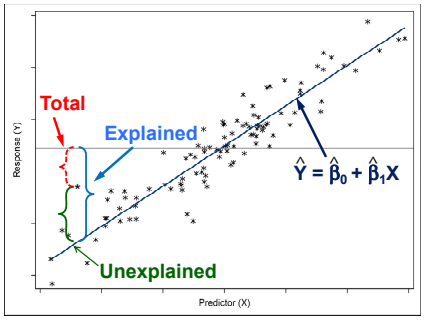

Figure 3: Explained vs. Unexplained variability.

A. Regression-Based Feature Selection

Fitting regression line with statistical measure of how the data is close called R-squared statistical measure or coefficient of determination. It employs a forward step-wise least squares regression that optimize the model r-squared value. It is used in huge data as a preparatory step to assess the features very fast and can identify quickly the useful features. The variations in target that explained by single feature with deleting the calculations of other features called squared correlation coefficient. The range of values between 0 and 1 (1 means the input feature can explain totally the variation in target) and 0 denotes that the target and input feature have not a relationship. In lung cancer recognition, the calculated squared correlation coefficient in a simple linear regression for all input features that are interval is mentioned as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{8}$$

where
X: input feature,
Y: response variable or target,
β_0: intercept parameter,
β_1: slope parameter
ε: error deviation of Y about $\beta_0 + \beta_1 X$.

The feature that explains the target is a worthy feature, thus it is selected in simple linear regression. In a baseline model, the target class and the input features have not a relationship. thus, any feature value does not improve predictions of the target class over simply using the mean of the target class for everyone.

R-Squared the ratio of variations explained via regression line in the observed data. The R-Squared is equal to $R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$, where SSM indicates sum square of model. It is the total variations explained by regression model and equal to $SSM = \sum(\hat{Y}_i - \bar{Y})^2$, SSE indicates to sum square of error. It is the total variations unexplained by regression model which means the error, and equal to $SSE = \sum(Y_i - \hat{Y}_i)^2$. Finally, SST indicates to total sum of square and equal to $SST = \sum(Y_i - \bar{Y})^2$. It is the correct total variations in the target class. Fig. 3 describes visually the relationships between the data, baseline model, total variability, explained variability and unexplained variability. A comparison between the squared correlation coefficient and the default Minimum R-Square of 0.005 is calculated and the feature is rejected if its value is

less than the cut-off criterion. The feature is selected if its R-square is greater than than the cut-off criterion. The sequential process of feature selection starts by choosing the feature that explains the large amount of changes in the target class. The stepwise process terminates when no remaining input feature can meet the Stop R-Square criterion.

B. Ensemble Based Feature Selection

Tree algorithms (decision trees, random forest, and gradient boosting) are powerful predictive models. They are the most widely used supervised learning due to their stability, high accuracy and ease of interpretation, considered to be one of the best and mostly used supervised learning methods. Decision tree is a type of supervised learning algorithm that is mostly used in classification problems. As shown in Fig. 4, a decision tree separates the data into segments, and a target value is assigned to each identical segment. A greedy, top-down recursive separating method is used. It employs exhaustive search at each phase by attempting all compositions of features and partition values to gain the maximum decrease in impurity. Subsequently, feature selection can be identified in tree building process. The process of selecting a specific feature based on its relative importance to split in impurity reduction can consider as a kind of feature selection. In our

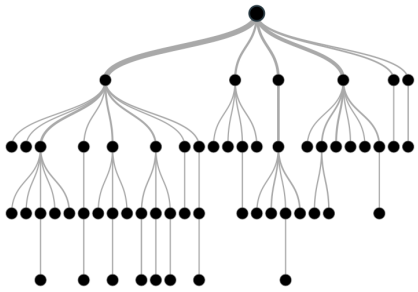

Figure 4: Decision Tree Diagram

model, the measures for feature importance or selection is based on the following metrics: count, surrogate count, residual sum square (RSS), and relative importance. The count-based feature importance simply counts the number of times in the tree that a particular feature is used in a split. Similarly, the surrogate count calculates the number of times that a variable is used in a surrogate splitting rule.

Feature importance measure is calculated for a single decision tree as:

$$VI(x_i, T) = \sum_{t \in T} \triangle I(x_i, t) \tag{9}$$

Where, $\triangle I(x_i, t) = I(t) - p_L I(t_L) - p_R I(t_R)$ is the reduction in impurity on feature x_i in tree T at a node t during split. p_L is the percentage of left observations by x_i and p_R for right. Gini index is calculated for classification of node t as:

$$Gini(t) = \sum_{i \neq j} p_i^t p_j^t \tag{10}$$

Where, p_i^t is the percentage of observations in t with class target equal i (y=i) and i,j run through target class. The Entropy= $-\sum_i p_i^t log(p_i^t)$ is similar to Gini index which evaluates impurity at a node t and its value is zero when node has observations from one class. When node has observations from mixture of classes, then entropy value is maximum.

Gradient boosting is a boosting approach that divides the dataset several times using random sampling to create outputs that form a weighted average of the re-sampled data set. Tree boosting generates a series of decision trees which together form a single predictive model. A tree in the series is fit to the residual of the prediction from the earlier trees in the series. The residual is defined in terms of the derivative of a loss function. For the stochastic tree ensemble (Gradient Boosting) of M trees, the generalized importance measure is calculated over the trees:

$$M(x_i) = \frac{1}{M} \sum_{j=1}^{M} VI(x_i, T_j) \tag{11}$$

In Gradient Boosting, separate models $f_k(x)$ are built to classify every k classes.

$$F_k(x) = \sum_{j=1}^{M} T_{kj}(x) \tag{12}$$

The general ((11)) is calculated as:

$$M(x_i, k) = \frac{1}{M} \sum_{j=1}^{M} VI(x_i, T_{kj}) \tag{13}$$

The total importance of x_i can be calculated with all classes as:

$$M(x_i) = \frac{1}{K} \sum_{k=1}^{K} M(x_i, k) \tag{14}$$

The proposed methodology for feature selection using gradient boosting is described in Algorithm 1 as:

VII. Support Vector Machine

Lung Cancer detection and recognition is formulated by binary classification problem. Each patient is classified as normal or abnormal. The goal is labeling a patient to detect the cancer and do the required steps. Many supervised learning methods are learned as computer aided system. In this section, we describe Support Vector Machines (SVMs) as an activity classifier we used in most of the experimental work presented in this field. SVMs are seen as relatively new supervised ML methodology developed by Cortes & Vapnik [14], which were first applied as an alternative to multi-layer neural networks.

begin
- Compute variable importance for a single decision tree as: $VI(x_i, T) = \sum_{t \in T} \triangle I(x_i, t)$
- Compute the reduction in impurity on feature x_i in tree T as: $\triangle I(x_i, t) = I(t) - p_L I(t_L) - p_R I(t_R)$
- For M trees, the generalized importance measure is: $M(x_i) = \frac{1}{M} \sum_{j=1}^{M} VI(x_i, T_j)$
- For every class, the generalized importance measure is: $M(x_i, k) = \frac{1}{M} \sum_{j=1}^{M} VI(x_i, T_{kj})$
- The total importance of x_i can be calculated with all classes as: $M(x_i) = \frac{1}{K} \sum_{k=1}^{K} M(x_i, k)$

end

Algorithm 1: PROPOSED GRADIENT BOOSTING FEATURE SELECTION ALGORITHM

To obtain the optimum decision boundary, SVM attempts to maximize the minimal distance from the decision boundary to the labeled data. Once this decision boundary is decided, a given unseen activity can be checked on which side of the decision boundary it lies (Fig. 5).

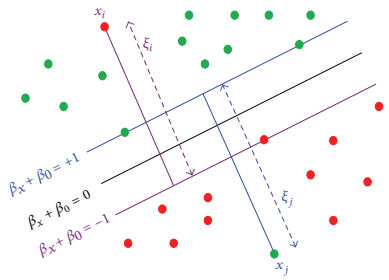

Figure 5: Support Vector Machine

Formally, let $\mathbf{S} = \{\{x_i, y_i\}_{i=1}^{n} \mid x_i \in \Re^d, y_i \in \{-1, +1\}\}$ be the training samples (i.e., feature vectors of patients), and $y_i \in \{-1, +1\}\}$ be the class label of x_i, thus two parallel separating hyperplanes can be formed such that:

$$y_i = \begin{cases} +1, & w^T x_i + b \geqslant 1 \\ -1, & w^T x_i + b \leqslant -1 \end{cases} \tag{15}$$

Where, T denotes the transpose operator, w is a perpendicular vector to the two hyperplanes and b is the bias, as shown in Fig. 5. Thus, the separating decision boundary (i.e. the optimal hyperplane) that maximizes the margin between the two classes is created by solving the following constrained optimization problem:

$$\begin{array}{ll} Minimize: & \frac{1}{2} \parallel w \parallel^2 \\ subject\ to & y_i(w^T x_i + b) \geqslant 1 \quad \forall_i \end{array} \tag{16}$$

By Lagrange duality, after some lengthy but straightforward calculations, the dual problem of the primal problem in (16) is given as:

$$\begin{array}{ll} Maximize: & \boldsymbol{W}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ subject\ to & \alpha_i \geqslant 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \end{array} \tag{17}$$

Where, $\alpha_i \geqslant 0$ are the lagrangian multipliers. Since (17) describes a Quadratic Programming (QP) problem, and a global maximum always exists for α_i, ω can be deduced as:

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{18}$$

VIII. SIMULATION RESULTS

The experiments are applied on kaggle DSB dataset. In this dataset, a thousand low-dose CT images from high-risk patients in DICOM format is given. The DSB database consists of 1397 CT scans and 248580 slices. Each scan contains a series with multiple axial slices of the chest cavity. Each scan has a variable number of 2D slices (Fig. 6), which can vary based on the machine taking the scan and patient. The DICOM files have a header that contains the necessary information about the patient ID, as well as scan parameters such as the slice thickness. It is publicly available in the Kaggle[1]. DICOM is the defacto file standard in medical imaging. This pixel size/coarseness of the scan differs from scan to scan (e.g. the distance between slices may differ), which can hurt performance of our model.

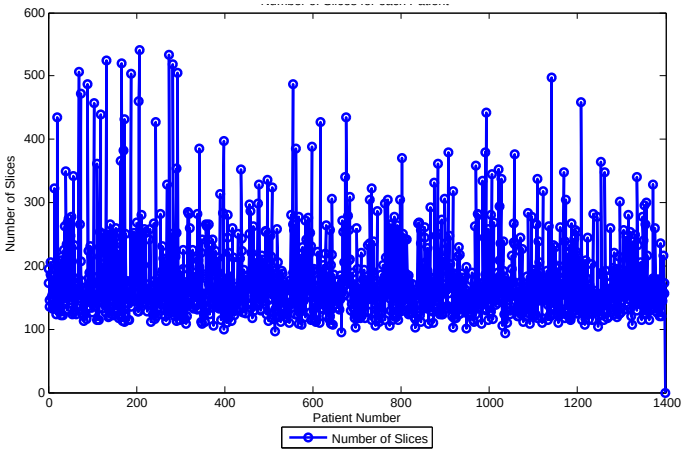

Figure 6: Number of slices per patient in data science bowl dataset.

The experiments are implemented on computer and its properties are described as follows: CPU i7, 2.6 GHz, 16 RAM, Matlab 2016b, R-Studio, and Python. Initially speaking, The nodules in Kaggle DSB dataset are detected and segmented using the watershed algorithm. The diameters of the nodules range from 3 to 30 mm. Each slice has 512×512 pixels and 4096 gray level values in Hounsfield Unit (HU), which is a measure of radiodensity.

After segmentation process, binarization process is done. In the screening setting, the annual low-dose CT study is one of the most difficult decisions whether CT or another investigation is needed. The nodule is very complex to be guided using current clinical guidelines due to its size and appearance. Moreover, the most important features of lung cancer are the size, number of nodules, location of the nodule,

[1]https://www.kaggle.com/c/data-science-bowl-2017/data

and type. The shape features are extracted which describe the shape of extracted nodules. Also, six moments are extracted for each slice of CT lung cancer that has a nodule. The concatenated features are selected based on forward step-wise least squared regression and gradient boosting to select the most important/relevant features and avoid the irrelevant, redundant features, and over-fitting problems.

Gradient Boosting builds a sequential decision tree to form a predictive model. Each iteration, the residuals of classification are updated using loss function derivatives from previous decision trees. The number of iterations in the boosting series is 50 iterations with 60% train proportion. In feature selection algorithm using R-square step-wise regression, the minimum R-square is 0.005 which is the cut-off threshold of a feature to be selected for R-square model selection and other features are irrelevant or redundant.

To evaluate the effectiveness of discrimination subset of features, we apply SVM with a linear kernel using 30% split for testing. Kaggle DSB dataset is divided into 60% for training, 10% for validation and 30% for testing. The accuracy of our proposed model is shown in Table I. As shown from Table I, the best accuracy is 88.07% with gradient boosting, watershed segmentation, and combination of moments and shape features.

Table I: Accuracy Results of our Algorithms on DSB Dataset

Method	Accuracy
Watershed+Moments+SVM	82.34%
Watershed+Shape+SVM	81.62%
Watershed+(Moments+Shape)+SVM	85.68%
Watershed+(Shape+Moments)+R2+SVM	87.11%
Watershed+(Shape+Moments)+Gradient Boosting+SVM	88.07%

The recognition results are shown by confusion matrix achieved on the DSB dataset with gradient boosting feature selection as shown in Table II. As shown from the Table II, Accuracy of model is 88.06%, Mis-classification rate is 11.93%, False positive rate is 11.29%, and False Negative is 13.76%. Almost all patients are classified correctly. Additionally, there is an enhancement on accuracy due to feature selection and efficient feature extraction.

Table II: Confusion Matrix of Watershed, Moments, Shape Feature, and Gradient Boosting Feature Selection using 30% Testing with SVM

	Predicted	
	Abnormal	Normal
Actual		
Abnormal	**94**	15
Normal	35	**275**

IX. Conclusion

Lung cancer recognition and detection based on gradient boosting and regression feature selection and watershed segmentation is presented. Due to the high dimensional data in medical images with hundreds of CT slices, feature selection is an important step to remove the irrelevant or redundant features. Also, the accuracy of models may be degraded with large number of features if there are not enough training observations to learn all parameters in model activities. In this paper, the features of CT scan for patients are extracted from simple and advanced discriminating method called shape and moments features, then feature selection methods are applied. The gradient boosting and regression models achieved the best accuracy compared to original features and state of the art. Also, ensemble-based or regression-based feature selection methods using random forest, gradient boosting, and R2 reduced the size of features which contribute to avoid over-fitting problems. In the future, we plan to investigate the problem of high dimensional data with different features, different datasets and different approaches like deep learning. Three dimensional convolution neural network (3D CNN) can improve the accuracy of model but it has more powerful machines to run with GPU.

References

[1] J. Schnabel and B. Glocker, "Deep learning for early detection of lung cancer in patients at risk," 2015.

[2] W.-J. Choi and T.-S. Choi, "Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach," *Entropy*, vol. 15, no. 2, pp. 507–523, 2013.

[3] S. S. Hamada R. H. Al-Absi, Brahim Belhaouari Samir, "A computer aided diagnosis system for lung cancer based on statistical and machine learning techniques," *JOURNAL OF COMPUTERS*, vol. 9, no. 2, pp. 425–431, 2014.

[4] B. B. Samir, K. B. Shaban, and S. Sulaiman, "Computer aided diagnosis system based on machine learning techniques for lung cancer," *International Conference on Computer & Information Science*, vol. 9, no. 2, pp. 295–300, 2012.

[5] M. Gomathi and P. Thangaraj, "A computer aided diagnosis system for lung cancer detection using support vector machine," *American Journal of Applied Sciences*, vol. 7, no. 12, pp. 1532–1538, 2010.

[6] Ada and R. Kaur, "Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images," *International Journal of Advanced Research in Computer Science and Softwere Enginering*, vol. 3, no. 3, pp. 187–190, 2013.

[7] S. L. Kumar, M. Swathy, S. Sathish, J. Sivaraman, and M. Rajasekar, "Identification of lung cancer cell using watershed segmentation on ct images," *Indian Journal of Science and Technology*, vol. 9, no. 1, 2016.

[8] K. Dimililer, B. Ugur, and Y. K. Ever, "Tumor detection on ct lung images using image enhancement," *The Online Journal of Science and Technology www.tojsat.*, vol. 7, no. 1, pp. 133–138, January 2017.

[9] N. S. Lingayat and M. R. Tarambale, "A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image," *International Journal of Bioscience, Biochemistry and Bioinformatics*, vol. 3, no. 6, pp. 624–629, 2013.

[10] D. Sharma and G. Jindal, "Computer Aided Diagnosis System for Detection of Lung Cancer in CT Scan Images," *International Journal of Computer and Electrical Engineering*, vol. 3, no. 5, pp. 714–718, 2011.

[11] A. Hashemi, A. H. Pilevar, and R. Rafeh, "Mass Detection in Lung CT Images Using Region Growing Segmentation and Decision Making Based on Fuzzy Inference System and Artificial Neural Network," *International Journal of Image, Graphics and Signal Processing*, vol. 5, no. 6, pp. 16–24, 2013.

[12] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, February 1962.

[13] J. Flusser and T. Suk, "Pattern recognition by affine moment invariants," *Pattern Recognition*, vol. 26, no. 1, pp. 167 – 174, 1993.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Journal of Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.