Clustering of lung adenocarcinomas classes using automated texture analysis on CT images
Antonio Pires¹, Henry Rusinek¹, James Suh², David P. Naidich¹, Harvey Pass³, Jane P. Ko¹
Departments of Radiology¹, Pathology² and Surgery³, New York University School of Medicine, NY, NY, 10002

ABSTRACT

Purpose: To assess whether automated texture analysis of CT images enables discrimination among pathologic classes of lung adenocarcinomas, and thus serves as an in vivo biomarker of lung cancer prognosis.

Materials and Methods: Chest CTs of 30 nodules in 30 patients with resected adenocarcinomas were evaluated by a pulmonary pathologist who classified each resected cancer according to the International Association for the Study of Lung Cancer (IASLC) system. The categories included adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), lepidic-predominant adenocarcinoma (LPA), and other invasive adenocarcinomas (INV). 3D volumes of interest (VOIs) and 2D regions of interest (ROIs) were then constructed for each nodule. A comprehensive set of N=279 texture parameters were computed for both 3D and 2D regions. Clustering and classification of these parameters were performed with linear discriminant analysis (LDA) using features determined by optimal subsets.

Results: Of the 30 adenocarcinomas, there were 13 INV, 11 LPA, 3 MIA, and 3 AIS. AIS and MIA groups were analyzed together. With all 3 classes, LDA classified 17 of 30 nodules correctly using the nearest neighbor (k=1) method. When only the two largest classes (INV and LPA) were used, 21 of 24 nodules were classified correctly. With 3 classes and 2D texture analysis, and when using only the two largest groups, LDA was able to correctly classify all nodules.

Conclusion: CT texture parameters determined by optimal subsets allows for effective clustering of adenocarcinoma classes. These results suggest the potential use of automated (or computer-assisted) CT image analysis to predict the invasive pathologic character of lung nodules. Our approach overcomes the limitations of current radiologic interpretation, such as subjectivity, inter- and intra-observer variability, and the effect of reader experience.

Keywords: Pulmonary nodule, lung cancer, adenocarcinoma, texture analysis, linear discriminant analysis, computed tomography, classification, computer-aided diagnosis

1. INTRODUCTION

Lung cancer is the most commonly diagnosed cancer worldwide, with 1.61 million cases reported in 2008.¹ Due to its high fatality rate, it is the leading cause of death from cancer.¹ Current CT scanners are able to detect pulmonary nodules as small as 1-2 mm in diameter and with each new generation of scanner, the ability to detect nodules improves.² Currently, guidelines recommend serial chest CT examinations to assess nodules for malignant behavior when solid, and at least three years when dealing if subsolid.²³⁴ A large portion of solid nodules are benign.² Subsolid nodules have faint “ground-glass” attenuation, less than that of soft tissue yet greater than adjacent lung attenuation. Subsolid nodules may have soft tissue components (part solid) or be pure “ground-glass”. These nodules have a high association with cancer when persistent, with 75% representing adenocarcinoma.⁵ Despite representing adenocarcinoma, many of these subsolid nodules have indolent behavior, thus subjecting patients to multiple CT scans and exposed to substantial amounts of radiation.⁶

Adenocarcinoma of the lung represents approximately 40% of all lung cancers and is the most common type of lung cancer among non-smokers.¹ There are several histological subtypes of adenocarcinoma, with adenocarcinoma comprised of mixed subtypes being the most common. However, recent changes in the classification system of adenocarcinoma now address clinical practice, research, and clinical trials.⁷ Adenocarcinoma in situ and minimally
invasive adenocarcinoma represent adenocarcinomas with purely lepidic growth or predominately lepidic growth with \( \leq 5 \) mm of invasion.\(^7\) These patients, if they undergo complete resection, will have near 100\% survival.\(^7\) Invasive adenocarcinomas are classified by histological pattern with lepidic adenocarcinomas now being comprised of mixed subtype tumors.\(^7\) These manifest as ground glass or part solid nodules, with 75\% representing adenocarcinoma or preneoplasia.\(^8\)

Patients with suspicious lesions undergo biopsy of the nodule, usually through minimally invasive procedures such as bronchoscope-guided aspiration, transthoracic CT-guided needle biopsy, or video-assisted thoracoscopic surgical (VATS) biopsy. All of these procedures are costly and entail potential complications. Moreover, knowledge gained in adenocarcinomas has created the need to examine the entire nodule pathologically to adequately characterize it histopathologically.\(^7\) For this reason in surgical candidates, biopsies of adenocarcinoma of the lung have less utility in diagnosis and staging of lung cancer given the heterogeneous nature. Therefore CT plays a major role in the current decisions regarding patient management with subsolid nodules that have a high likelihood of malignancy.

Radiologically, the malignant potential of pulmonary nodules can only be evaluated visually for features such as spiculated borders, air bronchograms, and cavitations.\(^9\) The novel methods of analysis include three-dimensional (3D) volume measurements of both solid and sub-solid components of nodules to assess growth rate in the context of diagnosing whether or not a nodule is malignant.\(^10\) The use of computer-aided analysis of CT texture features allows for the potential classification of lung nodules and mainly have addressed solid nodules and margin analysis.\(^11\) Such textures features would aid in differentiating lesions of greater aggressiveness histopathologically, and therefore worse prognosis, for which more aggressive management would be indicated such as resection, from more indolent adenocarcinomas.\(^12\) However, no information on correlation with the type of malignancy is available. Texture analysis, a type of computer image analysis, allows an image (computed tomography images in our case) to be quantified in terms of organization, intensity, spatial relationships, etc.\(^13\)

Here, we present the use of such texture parameters used to distinguish between 3 categories of adenocarcinoma.

### 2. METHODS

#### 2.1 Source images and regions of interest (ROIs)

Chest CTs images (high-frequency kernel, 0.8-1.25 mm overlapping sections) of 30 nodules in 30 patients with resected adenocarcinomas had been evaluated by a pulmonary pathologist who classified each resected cancer according to the IASLC system as adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), lepidic-predominant adenocarcinoma (LPA), and other invasive adenocarcinomas (INV). CT scans were performed using 120 kVp, tube current modulation, reference values of 80 to 120 mAs with reconstructions using 35-40 cm standard field of view. Image data was anonymized prior to evaluation.

Images were converted from DICOM to 3D raw format. Using Mazda v3.3 and Mazda v4.6 (Technical University of Lodz, Poland), 3D volumes of interest (VOIs) were then constructed for each nodule using a segmentation technique based on connectivity, compactness, and thresholding (Figure 1).\(^14\)\(^-\)\(^16\) 2D ROIs were also created using an electronic paintbrush on the CT slice with the largest area of nodule (Figure 1).

#### 2.2 Texture features

A comprehensive set of \( N=279 \) texture parameters were computed for both 3D and 2D regions using Mazda v3.3 and Mazda v4.6.\(^9\)\(^-\)\(^11\) The features are derived from the histograms of CT attenuation (\( N=9 \)), histogram of the magnitude of image gradient (\( N=5 \)), run-length matrix characteristics (\( N=20 \)), co-occurrence matrix features (\( N=220 \)), autoregressive features (\( N=5 \)), and Haar wavelet coefficients (\( N=20 \)).

Some of the most commonly selected texture parameters in our analysis of the co-occurrence matrix included entropy, correlation, difference entropy, sum entropy, and inverse difference moment. Other common features not associated with
the co-occurrence matrix came from the run-length matrix such as the horizontal long run emphasis and 45-degree fraction, which will also be discussed.

2.2.1 Co-occurrence matrix features

The co-occurrence matrix, $P[i,j]$, is defined by first specifying a displacement vector $d = (dx,dy)$. The size of the matrix $P[i,j]$ is defined by the number of gray levels, so that an image with $N$ gray levels will have an $N \times N$ co-occurrence matrix associated with it. Each pair of pixels satisfying the separation given by the displacement vector constitutes one entry in $P[i,j]$ where $i$ represents the gray level of the first pixel and $j$ represents the gray level of the second pixel.

The Mazda program used 20 vectors: $(x,0)$, $(0,x)$, $(x,x)$, and $(x,-x)$ with $x$ ranging from 1 to 5 to compute a normalized co-occurrence matrix for each vector. For each co-occurrence matrix, a number of texture features were calculated, the most relevant of which are explained below.

2.2.1a Entropy

The entropy of an image, defined in Eq. 1, refers to the randomness of the distribution of gray levels such that an image with high entropy has no preferred set of gray level pairs, and as a result, has a uniformly populated co-occurrence matrix.

$$\text{Entropy} = -\sum_i \sum_j P[i,j] \log P[i,j]$$

Visually, an image with high entropy would consist of randomly dispersed gray levels with no particular order, and as such, entropy is a useful measure of disorganization. Rapidly growing tumors may be more disorganized and therefore have less structure, which means higher entropy.

2.2.1b Contrast and inverse difference moment

Both of these measures in Eqs. 2 & 3 look at the gray level difference between pixel pairs, but contrast increases for larger differences between gray level pairs whereas the inverse difference moment (IDM, also known as homogeneity) decreases for larger differences. Contrast increases as the image becomes more inhomogeneous and therefore proves useful in images with several contrasting components. For example, with regards to high-resolution CT of lung cancer, contrast is higher (IDM is lower) in the invasive tumors containing solid tumor tissue, which would contrast with the ground glass components of the surrounding lepidic tissue. Also, contrast and IDM may be affected by blood vessels contrasted against different types of tissue within the tumor. Furthermore, image contrast and IDM would increase with the administration of intravenous contrast.

$$\text{Contrast} = \sum_i \sum_j (i-j)^2 P[i,j]$$

$$\text{IDM} = \sum_i \sum_j \frac{1}{1+(i-j)^2} P[i,j]$$

2.2.1c Sum entropy and difference entropy

These two measures are calculated similarly to entropy but use values based on the sums or differences of gray levels pairs as defined in Eqs. 4 & 5 below:

$$P_{i+j}(k) = \sum_{i+j=k} P[i,j] \quad k = 0,1,2...,2K-2$$
\[ P_{i-j}(k) = \sum_{|i-j|=k} P[i,j] \quad k = 0,1,2,...,K - 1 \]  

That is for every possible sum or difference of gray levels, \( P_{i,j} \) and \( P_{i,j} \) give the probability of having such a pair from the image for each \( k \). These two functions can then be used to calculate texture parameters such as the sum entropy and difference entropy as follows:

\[
\text{Sum Entropy} = -\sum_k P_{i+j}(k) \log(P_{i+j}(k))
\]

\[
\text{Diff Entropy} = -\sum_k P_{i-j}(k) \log(P_{i-j}(k))
\]

While it is difficult to visually interpret these two parameters in Eqs. 6 & 7, they represent the same type of parameter that entropy does in that it measures disorganization using gray level sums and differences rather than simply gray levels.

### 2.2.2 Run-length matrix features

Briefly, the run-length matrix is defined as \( P_k[i,j] \), which counts the number of runs (strings of adjacent pixels with the same gray level) having length \( j \) with a gray level of \( i \) at angle \( k \). Mazda calculated a run-length matrix for angles of 0, 45, 90, and 135.

#### 2.2.2a Horizontal long run emphasis

A commonly encountered feature in our study was the horizontal long run emphasis (HRLE) defined in Eq. 8, where \( n_r \) represents the total number of runs at angle of 0 degrees. It is weighted towards longer runs.

\[
HLRE = \frac{1}{n_r} \sum_i \sum_j P_0[i,j] \cdot j^2
\]

#### 2.2.2b Fraction of image in runs at 45 degrees (45Frac)

Another run-length matrix feature is the fraction of the image in runs, Eq. 9. It measures the fraction of image pixels that are part of the runs considered during the run-length matrix computation.

\[
45Frac = \frac{\sum_i \sum_j P_{45[i,j]} \cdot j \cdot P_{45[i,j]}}{\sum_i \sum_j P_{45[i,j]}}
\]

HLRE increases for smoother or more uniform images, as these types of images will have longer runs of pixels with similar gray levels. Histologically it may represent amounts of similar tissue within lung nodules. 45Frac may be different for different types of tissue found in lung nodules. For example, if one type of tissue in a nodule spreads out more than another type of tissue found in another type of nodule, then they will have different run-length matrices and more runs will be included with the nodule containing the larger area of the specific tissue, assuming that the tissue is relatively homogeneous.

### 2.3 Clustering and classification

Feature reduction was achieved using the optimal subsets method with the option of using feature triplets to determine the features with the lowest classification error. Linear discriminant analysis (LDA) used the features selected by the optimal subsets method to cluster each of the 30 nodules using transformations of the original data to create \( N-1 \)
variables called most discriminating features (MDF), where \( N \) is the number of nodule groups. The data is then classified based on the 1-nearest neighbor method. (Refer to Mazda manual for in depth discussion of feature reduction and classification methods: http://www.eletel.p.lodz.pl/programy/mazda/index.php?action=docs.)

3. RESULTS

Overall, of the 30 adenocarcinomas, there were 11 LPA, 13 INV, 3 MIA, and 3 AIS. The AIS and MIA groups were analyzed together given the small size of each group and similar 5-year survival. Through data analysis, the optimal subsets method using feature triplets proved the best for identifying discriminative features to use for LDA.

With all 3 classes, optimal subsets determined 10 features that were discriminative for the 3D VOI texture approach. LDA allowed 17 of 30 nodules to be correctly classified using the 1-NN method (Figure 2). When only the two largest classes (INV and LPA) were used, 21 of 24 nodules were classified correctly (Figure 3) using 15 features selected by optimal subsets.

With 3 classes and 2D texture approach, 21 texture features were selected by the optimal subsets method. With LDA, the 1-NN method was able to classify 30 of 30 nodules correctly (Figure 4). When using only the two largest groups (invasive and lepidic), optimal subsets determined that there were only 12 discriminative texture features and LDA was able to classify 24 of 24 nodules correctly (Figure 5).

4. DISCUSSION

Texture parameters determined by optimal subsets using triplet features allows for the clustering of adenocarcinoma classes using CT images and LDA. These preliminary results suggest the potential use of CT to predict the invasive pathologic character of lung nodules using image texture features. Surprisingly, 3D features appeared less reliable than using 2D features. This may be due to the effect of variable reconstructed slice thickness (0.8-1.25 mm) and less accurate nodule segmentation in a 3D approach given the ground-glass components and manual technique for segmentation.

Limitations of the current work are the small sample size, as the smallest category of nodules only contained 6 nodules. Another drawback is that due to limitation of texture analysis software, ROIs and VOIs were drawn by hand, which introduces variability depending on who creates the ROI or VOI.

In spite of these drawbacks, the accurate classification results are highly encouraging and call for future confirmatory studies. Texture based classification approach overcomes the limitations of current radiologic interpretation such as subjectivity, inter- and intra-observer variability, and the effect of reader experience. With this approach, a rich set of imaging features and the interaction among features are accounted for, which may yield additional information compared to analyzing individual features.17

REFERENCES

Figure 1: The top row displays a single 2D CT image of a nodule before and after the creation of an ROI in Mazda. The lower row of images displays the nodule in a 3D format before and after the creation of a VOI in Mazda.
Figure 2: This graph displays the results of LDA for 3D VOI texture approach using all categories. Classification with 1-NN classified 17 of 30 nodules correctly. Note that there are N-1 MDFs. 1=LPA, 2=AIS/MIA, 3=INV.

Figure 3: This graph displays the results of LDA for 3D VOI texture approach using the two largest categories. Classification with 1-NN classified 21 of 24 nodules correctly. 1=LPA, 2=INV.
Figure 4: This graph displays the results of LDA for 2D ROI texture approach using all categories. Classification with 1-NN classified 30 of 30 nodules correctly. 1=LPA, 2=AIS/MIA, 3=INV.

Figure 5: This graph displays the results of LDA for 2D ROI texture approach using two categories. Classification with 1-NN classified 24 of 24 nodules correctly. 1=LPA, 2=INV.