10-1-2010

Predicting Panel Ratings for Semantic Characteristics of Lung Nodules

Dmitriy Zinovev
DePaul University

Jonathan Feigenbaum
DePaul University

Daniela Raicu
DePaul University

Jacob Furst
DePaul University

Predicting Panel Ratings for Semantic Characteristics of Lung Nodules

Dmitriy Zinovev
DePaul University, CDM
243 S. Wabash Ave,
Chicago IL. 60604
1(630)639-3417
dzinovev@gmail.com

Jonathan Feigenbaum
DePaul University, CDM
243 S. Wabash Ave,
Chicago IL. 60604
1(847)894-1602
jfeigenbaum@bus.illinois.edu

Daniela Raicu
DePaul University, CDM
243 S. Wabash Ave,
Chicago IL. 60604
1(312)362-5512
dstan@cdm.depaul.edu

Jacob Furst
DePaul University, CDM
243 S. Wabash Ave,
Chicago IL. 60604
1(312)362-5158
jfurst@cdm.depaul.edu

ABSTRACT

In reading CT scans with potentially malignant lung nodules, radiologists make use of high level information (*semantic characteristics*) in their analysis. CAD systems can assist radiologists by offering a "second opinion" - predicting these semantic characteristics for lung nodules. In our previous work, we developed such a CAD system, training and testing it on the publicly available Lung Image Database Consortium (LIDC) dataset, which includes semantic annotations by up to four human radiologists for every nodule. However, due to the lack of ground truth and the uncertainty in the dataset, each nodule was viewed as four distinct instances when training the classifier. In this work, we propose a way of predicting the distribution of opinions of the four radiologists using a multiple-label classification algorithm based on belief decision trees. We evaluate our results using a distance-threshold curve and, measuring the area under this curve, obtain 69% accuracy on the testing subset. We conclude that multiple-label classification algorithms are an appropriate method of representing the diagnoses of multiple radiologists on lung CT scans when a single ground truth is not available.

1. INTRODUCTION

Lung cancer is the most prevalent cause of cancer-related deaths in the human population today. Effective treatment often relies on early detection of the disease, which is done by analyzing suspect computed tomography (CT) scans of lungs. Analysis of the size change of suspected tumors – known as lung nodules – and the inspection of their visual characteristics helps to diagnose the patient.

To increase the likelihood of correct diagnosis by a human radiologist, computer-aided diagnosis (CAD) systems have been designed. These algorithms are classifiers trained on pre-existing data and then used to classify new CT scans. They provide a "second opinion" to the radiologist, which may help to increase the efficiency of the diagnosis process as well as reduce the rate of false positive diagnoses while maintaining an acceptably low rate of false negative diagnoses at the same time.

These algorithms are trained and tested on different datasets of lung CT scans. One such dataset is the Lung Image Database Consortium (LIDC) – a diverse and growing collection of CT scans analyzed by four radiologists. Each radiologist provided a contour for the nodule or nodules present in the scan, as well as a set of characteristics for the nodule as a whole (cross sections of the same nodule are generally present on multiple CT scans). These characteristics are lobulation, malignancy, margin, sphericity, spiculation, subtlety and texture. Each characteristic received a rating on a scale from one to five.

The LIDC provides a common framework for training and evaluating CAD algorithms, which allows the effectiveness of different algorithms to be compared more easily than if proprietary data were used. However, there are certain disadvantages to the dataset; the two that we are most concerned with are the

lack of ground truth and the disagreement between multiple observers – there was no forced consensus among radiologists in their ratings. The only nodules that can be considered marked reliably are those on which all (or most) radiologists agreed with respect to outline and semantics. Unfortunately, the four radiologists who looked at each scan often disagreed about both the outline of the nodule and the characteristic ratings for the nodule as a whole. These factors complicate the training of CAD systems, since it is unclear how the reference truth [1] is to be derived and the results of classification evaluated.

In our previous work, we built a CAD algorithm that classified the characteristics of each nodule based on 63 image features that can be calculated from the scan of the nodule. These features were then used to predict the ratings for each of the seven semantic characteristics, thus providing a second opinion to a radiologist who needs to classify a nodule. Since there were four radiologists, they provided four sets of outlines and four sets of ratings for each nodule. Different outlines produce different image features, so there was a total of four sets of features corresponding to four sets of ratings. We addressed this ambiguity by considering every set of features and corresponding ratings as a separate case in training the algorithm. Another possible approach for addressing disagreement issues before training the classification model is to artificially force a diagnosis consensus by finding the mean or mode of the features and characteristics. The potential drawback of such solution is a loss of important information in some cases. For instance, if two radiologists rated a nodule as a one on the malignancy scale (benign) and two others marked it as a five (malignant), the average would be a three (uncertain). In this way, a case deserving additional attention would risk getting dismissed. Also the fact of disagreement can itself be used as information characterizing the nodule.

In this paper, we took a further step of classifying the nodules of LIDC dataset by combining the data from the four radiologists without losing important information that may be present in the distribution of ratings. We consolidated the four sets of characteristics available for each nodule into a single distribution, where each rating (from one to five) received a frequency based on the proportion of radiologists who selected that rating. The single set of image features was calculated for the largest of the four outlines provided by the radiologists, as it was considered the most representative. The algorithm was then trained on the set of image features and the distribution of characteristic ratings for each nodule. In this way, we were able to represent each nodule as a single case during the training while taking full advantage of all the information available in the radiologists' ratings, instead of discarding the deviations.

2. RELATED WORK

In the classification task, a classification instance is a case that has to be assigned a label, a label is a class membership/membership probability of a particular instance and class defines the group that the instance can be a member of. Sometimes a label for a classification instance might assign that particular instance to a multiple classes. All tasks of classifying such instances can be divided into two distinct groups: multi-label and multiple-label classification. There are also a family of classification problems that deal with instances that are associated with only one class label, but the class is chosen from a pool of classes where number of classes is greater than two. Work of Hu et al [2] defines such task as multi-class problem and proposes a solution for it employing Support Vector. Finally, classification problems that deal with instances defined by multiple sets of attributes are called multi-instance. A solution for this class of problems was proposed by Dietterich et al [3]. Table 1 summarizes these techniques in the context of the LIDC dataset.

Table 1. Summary of different classification tasks

Task	Definition	Example	Relation to LIDC
Multi-label	Each classification instance could be a member of several independent classes simultaneously.	Scene classification on images. Image can be a member of classes beach, forest and mountains simultaneously.	Cannot be applied to classification of nodules contained in LIDC dataset since classes (ratings) are not independent and instance cannot be a true member of multiple classes simultaneously.
Multiple-label	While classification instance can have a multiple class labels associated with it (for example due to the multiple observers), only one of them could be correct.	Imagine that the vehicle engine is diagnosed by two different diagnosis tools and one of them (not known which one) is defective. They will produce different diagnosis with only one of them being correct.	Partially to classification of nodules contained in LIDC dataset. Instance can have a multiple class labels associated with it and usually only one of them is correct. However, it is possible that for some instance none of class labels associated with it are correct.
Multi-class	The classification instance can have only one class label associated with it, but the class is chosen from a pool of classes where number of classes is greater than two.	Automatic traffic analyzer. The program has to assign recognized vehicle to one of existing classes {car, truck, bus, motorcycle}	Cannot be applied to classification of nodules contained in LIDC dataset. Due to the presence of multiple observers, the instance can have several class labels associated with it.
Multi-instance	Each classification instance can have multiple feature vectors associated with it (again, one of the possible reasons – presence of multiple observers), yet only one feature vector is the one responsible for true class of the instance.	Classification of good product vs. bad product where characteristics of a good product are obtained doing a survey in a group of customers	Partially to classification of nodules contained in LIDC dataset. Instance can have multiple feature vectors associated with it (due to the multiple outlines), but we are not guaranteed that any of them is responsible for true class of the instance, because instance can not have it's true class associated with it

The multi-label classification task is applicable in the situations when the instance can be a member of several non mutually exclusive classes simultaneously. The examples of such tasks (of video, gene and image classification) are described in [4,5,6]. In addition, Shen et al. [7] describe the classification of scene pictures with respect to the content. The paper defines a number of classes that an image can belong to {Beach, Sunset, Foliage, Field, Mountain, Urban}. A separate classifier is trained for each of the classes, and the decision that classifier makes is binary, i.e. the instance is either a member or non-member of the particular class. This is achieved either by use of a classifier that is binary by nature or by thresholding [7] if the output of the classifier is a probability.

Another example of a multi-label classification task is the classification of a movie genre. In this case possible classes that a movie can belong to are the different genres {comedy, drama, action, documentary, horror, romance, independent}. Constructed classifiers have to decide whether the movie belongs to the comedy genre or not, whether it belongs to the drama genre or not, etc. Such a task and a solution for it were described by Veloso et al. in [9].

The multiple-label classification task defined by Jin et al. [10] is similar to the multi-label classification task in a sense that the instance can be a member of several classes at the same time, but differs from it by the fact that only one of these class memberships is correct. Situations in which such a classification task is applicable usually arise due to the presence of multiple observers who do not agree with each other. There are two possible approaches for solving such classification tasks: problem transformation and algorithm adaptation approaches.

In the problem transformation approach the learning task is transformed in such a way that it becomes suitable for standard single-label classification techniques. Tsoumakas et al [11] described several

problem transformation techniques including the “copy” and “select” families of techniques, and the label powerset and binary relevance techniques.

The algorithm adaptation approach assumes modification of some existing classification technique in such a way that would make the technique able to handle instances with labels being indicators of multiple class memberships. The example of use of such approach was described by Bjanger and Denœux [12] in which they modified a regular decision tree algorithm by defining the impurity measure for each node with respect to a class membership probability distribution of an instance as opposed to single class membership. Another approach for solving the multiple-label problem using artificial Neural Networks was proposed in a work of Denœux [13] and in the work of Quost and Denœux [14] where authors employed the Dempster Shaefer theory [15] for combining uncertain output labels produced by multiple weak classifiers for identifying different types of waveforms in sleep EEG data. The dataset was obtained by collecting 64 measurements of brain activity separated by 2 second interval. Two possible classes for the collected brain waves were K-complex and delta wave-forms. Signals were manually identified by 3 experts and due to the complexity of identifying the K-complexes, experts did not always agree therefore producing uncertainty in the labels. The approach proposed by the authors demonstrated an error rate of 13.4. For every classification case the authors performed minimization of mean squared differences between the classifier outputs and target values making a decision on whether the instance was classified correctly or not.

Bjanger and Denœux proposed an adaptation of decision trees classifier to the multiple-label problem, first for classifying uncertain two-class label instances for classifying the EEG data (classification approach has shown error rate of 0.34) [12] and later (by Vannoorenberghe and Denœux) uncertain multiple class label instances [16] by combining trees produced by splitting single multiple-label classification problem into multiple two-class classification problems (one vs. rest classification approach) for classifying data concerning acoustic emission testing of pressure vessels. The dataset consisted of clusters of acoustic emission signals, each cluster belonging to one of three classes. Each training example was accessed by two experts who were asked to assign the degree of possibility that the example belongs to each class. This resulted in 2 possibility distributions per case. The uncertain labels were constructed either by taking into account each expert’s opinion individually or by combining two opinions (or decisions made by two individual classifiers). The reported results had shown 0.32, 0.29, 0.29 and 0.3 error rates correspondingly. The output of such classifier is another basic belief assignment (BBA) that can be evaluated against the original uncertain label using various metrics such as simple accuracy as in [15] or loss function proposed at [16].

When classification is applied to the task of classifying the lung nodules contained in the LIDC dataset [17] possible classes are ratings {1-5} that can be assigned to semantic characteristics of a nodule. For most of the semantic characteristic the ratings are ordinal. Each nodule is annotated and outlined by up to four different radiologists who do not necessarily agree with each other, therefore producing up to 4 different sets of ratings and 4 different outlines per nodule. Even though the nodule might have several ratings per semantic characteristic associated with it, it is obvious that only one of them is correct (a nodule cannot be malignant and benign at the same time). This description clearly defines the problem of classifying the lung nodules contained in the LIDC dataset as a multiple-label classification problem. When classifying a single nodule we are solving seven independent multiple-label classification problems, one for every semantic characteristic. The described LIDC nodules classification problem fits the definition given by Jin [10] with the assumption that at least one of the radiologists provides the correct label corresponding to the ground truth.

In our previous work we built a classification model that predicted a rating assigned to a nodule by a single radiologist. The set of attributes for the instance was calculated from the outline provided for this nodule by the given radiologist. Given the fact that the nodule could have up to 4 different ratings and outlines associated with it the problem of LIDC nodules classification fits the definition of multiple-label, multi-instance classification. By making prediction on a level of single interpretation/outline we reduced the task to a common multi-class classification problem. . In this paper we will describe an approach capable of solving a multiple label classification task in the context of the LIDC dataset. We will present an algorithm capable of predicting a class membership probability distribution for a nodule and an evaluation metric capable of assessing the performance of such an algorithm. Table 2 gives a summary of differences between the previous and the current work. The rest of the paper is organized as follows: we describe the classification task in detail along with the description of a classification approach in the "methodology" section, present the classification results and their evaluation in the "results" section and describe our plans for future work in the"conclusion" section.

Table 2. Detailed comparison of previous and current work.

	Previous work	**Current work**
Label	Single label (Radiologist's rating)	Multiple label (distribution of radiologist ratings)
Dataset	Number of instances = number of nodules * up to 4 interpretations	Number of instances = number of nodules
Features	Attributes ←largest outline provided by given radiologist for given nodule	Attributes ←largest outline provided for given nodule
Evaluation	Assigned rating vs. mode of predicted multiple-label	Assigned multiple-label vs. predicted multiple-label

3. METHODOLOGY

3.1 LIDC Dataset

The publicly available LIDC database [17] (downloadable through the National Cancer Institute's Imaging Archive web site - http://ncia.nci.nih.gov/) provides the image data, the radiologists' nodule outlines, and the radiologists' subjective ratings of nodule characteristics for this study. The LIDC database currently contains complete thoracic CT scans for 208 patients acquired over different periods of time and with various scanners.

The XML files accompanying the LIDC DICOM images contain the spatial locations of three types of lesions (nodules < 3 mm in maximum diameter, but only if not clearly benign; nodules > 3 mm but < 30 mm regardless of presumed histology; and non-nodules > 3 mm) as marked by a panel of 4 LIDC radiologists. For any lesion marked as a nodule > 3 mm, the XML file contains the coordinates of nodule

outlines constructed by any of the 4 LIDC radiologists who identified that structure as a nodule > 3 mm. Moreover, any LIDC radiologist who identified a structure as a nodule > 3 mm also provided subjective ratings for 9 nodule characteristics subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, and malignancy likelihood. For a detailed description of semantic characteristics, refer to table 3.

Table 3. LIDC nodule characteristics with corresponding rating scale

Characteristic	Notes and References	Possible Scores
Calcification	Pattern of calcification present in the nodule	1. Popcorn 2. Laminated 3. Solid 4. Non-central 5. Central 6. Absent
Internal structure	Expected internal composition of the nodule	1. Soft Tissue 2. Fluid 3. Fat 4. Air
Lobulation	Whether a lobular shape is apparent from the margin or not	1. Marked 2. . 3. . 4. . 5. None
Malignancy	Likelihood of malignancy of the nodule - Malignancy is associated with large nodule size while small nodules are more likely to be benign. Most malignant nodules are non-calcified and have spiculated margins.	1.Highly Unlikely 2.Moderately Unlikely 3. Indeterminate 4.Moderately Suspicious 5. Highly Suspicious
Margin	How well defined the margins of the nodule are	1. Poorly Defined 2. . 3. . 4. . 5. Sharp
Sphericity	Dimensional shape of nodule in terms of its roundness	1. Linear 2. . 3. Ovoid 4. . 5. Round
Spiculation	Degree to which the nodule exhibits spicules, spike-like structures, along its border - Spiculated margin is an indication of malignancy	1. Marked 2. . 3. . 4. . 5. None

Subtlety	Difficulty in detection - Subtlety refers to the contrast between the lung nodule and its surrounding	1. Extremely Subtle 2. Moderately Subtle 3. Fairly Subtle 4.Moderately Obvious 5. Obvious
Texture	Internal density of the nodule - Texture plays an important role when attempting to segment a nodule, since part-solid and non-solid texture can increase the difficulty of defining the nodule boundary	1. Non-Solid 2. . 3. Part Solid/(Mixed) 4. . 5. Solid

The LIDC dataset contains a number of CT studies performed on a number of patients over an extensive period of time. Each study can contain several nodules of a different size; therefore, there may be a different number of slices associated with a particular nodule. Each slice associated with a nodule could contain up to 4 different outlines of this nodule marked by 4 different radiologists. Each radiologist independently rates 7 semantic characteristics of a nodule which produces 4 different semantic labels associated with it (Figure 1). Since radiologists often disagree on the existence of a nodule in a particular location, there might be less than 4 different semantic labels associated with given nodule; therefore, the description of LIDC nodules is not uniform.

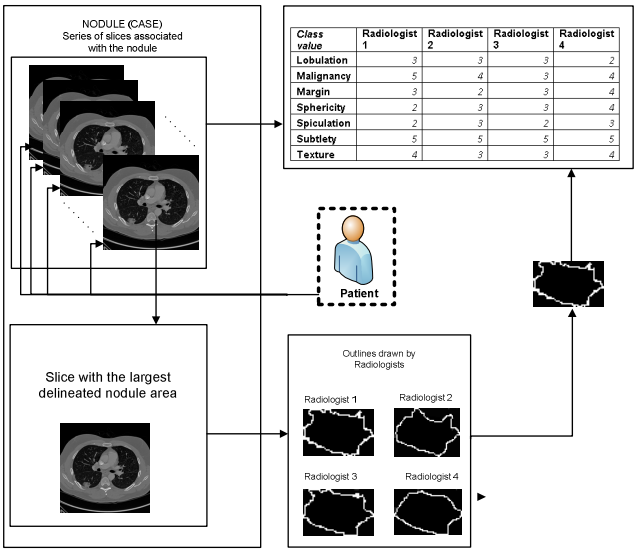


Figure 1. Visual representation of multi-label multi-class classification task in the context of theLIDC data

3.2 Image Feature Extraction

For each nodule greater than 5×5 pixels (around 3×3 mm) - nodules smaller than this would not have yielded meaningful texture data – we calculate a set of 63 two-dimensional (2D), low-level image features from four categories: shape features, texture features, intensity features, and size. Although each nodule is present in a sequence of slices, in this study we are considering only the slice in which the nodule has the largest area along with up to four (depending on the number of radiologists detecting and annotating the corresponding nodule) image instances corresponding to this slice. After completion of the feature extraction process, we created a vector representation of every nodule image which consisted of 63 image features and 9 radiologist annotations.

Size Features - We use the following seven features to quantify the size of the nodules: area, ConvexArea, perimeter, ConvexPerimeter, EquivDiameter, MajorAxisLength, and MinorAxisLength. The area and perimeter image features measure the actual number of pixels in the region and on the boundary, respectively. The ConvexArea and ConvexPerimeter measure the number of pixels in the convex hull and on the boundary of the convex hull corresponding to the nodule region. EquivDiameter is the diameter of a circle with the same area as the region. Lastly, the MajorAxisLength and MinorAxisLength give the length (in pixels) of the major and minor axes of the ellipse that has the same normalized second central moments as the region.

Shape Features - We use seven common image shape features: circularity, roughness, elongation, compactness, eccentricity, extent, and the standard deviation of the radial distance. Circularity is measured by dividing the circumference of the equivalent area circle by the actual perimeter of the nodule. Roughness can be measured by dividing the perimeter of the region by the convex perimeter. A smooth convex object, such as a perfect circle, will have a roughness of 1.0. The eccentricity is obtained using the ellipse that has the same second-moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 (a perfect circle) and 1 (a line). Solidity is the proportion of the pixels in the convex hull of the region to the pixels in the intersection of the convex hull and the region. Extent is the proportion of the pixels in the bounding box (the smallest rectangle containing the region) that are also in the region. Finally, the RadialDistanceSD is the standard deviation of the distances from every boundary pixel to the centroid of the region.

Intensity Features - Gray-level intensity features used in this study are simply the minimum, maximum, mean, and standard deviation of the gray-level intensity of every pixel in each segmented nodule and the same four values for every background pixel in the bounding box containing each segmented nodule. Another feature, IntensityDifference, is the absolute value of the difference between the mean of the gray-level intensity of the segmented nodule and the mean of the gray-level intensity of its background.

Texture Features - Normally texture analysis can be grouped into four categories: model-based, statistical-based, structural-based, and transform-based methods. Structural approaches seek to understand the hierarchal structure of the image, while statistical methods describe the image using pure numerical analysis of pixel intensity values. Transform approaches generally perform some kind of modification to the image, obtaining a new “response” image that is then analyzed as a representative proxy for the original image. Model-based methods are based on the concept of predicting pixel values based on a mathematical model. In this research we focus on three well-known texture analysis techniques: co-occurrence matrices (a statistical-based method), Gabor filters (a transform-based method), and Markov Random Fields (a model based method).

3.3 Adaptation of Belief Decision Trees for LIDC dataset

In this paper we chose to adopt the classification approach proposed by Elouedi [18]. It is a version of decision trees that is able to handle data instances with uncertain labels. Classification is performed in a manner similar to the one of regular decision trees. On every node, the instance that is currently being classified is redirected to the right or the left child of the node depending on the value of the attribute corresponding to this node. The process is repeated until the instance reaches the leaf node, which has a class membership probability distribution or a basic belief assignment (BBA) associated with it. This BBA is considered to be the newly predicted label of a classified instance. The main difference lies in the way a tree is constructed. At every node of the tree, starting with the root, the algorithm attempts to perform a split based on every attribute/feature existing in the dataset. Out of all constructed splits it determines the best (the selection measure will be defined further) one and uses it for growing the tree further. Every node is associated with a BBA that is constructed by the average of the BBAs of all training cases that reached that node. The newly created node is considered to be a leaf if one of the stopping criteria is reached: 1) there is only one instance that reached this node; 2) all BBAs of the instances which reached the node are equal; 3) all the available attributes/features are split; or 4) the gain ratio of all possible further splits is less than or equal to 0.

In order to define a best split, the algorithm performs the following steps:

Computes the pignistic probability (probability calculated from a belief) of instance I_j for each possible class C_i for every instance in the dataset by:

$$BetP^{\Theta}\{I_j\}\{C_i\} = \sum_{C_i \in C \subseteq \Theta} \frac{1}{|C|} \frac{m^{\Theta}\{I_J\}(C)}{1 - m^{\Theta}\{I_J\}(0)}, \forall C_i \in \Theta \quad (1)$$

Where C is a belief mass that C_i is a member of, Θ is a set of all possible classes and $m^{\Theta}\{I_J\}(C)$ is a probability associated with the corresponding belief mass C. Due to the fact that all BBAs in the LIDC dataset are singletons, the pignistic probability of instance I_j for class C_i is the ratio of observers who assigned the instance to a given class to the total number of observers for that instance (equation 2).

$$\lambda_{ratl} = \frac{\lambda_l}{\sum_{i=1}^{5} \lambda_i} \quad (2)$$

(where $\lambda_l=\{0,1,2,3,4\}$ is rater count for every class l rated on a scale from 1 to 5

Next, the algorithm computes the average pignistic probability function $BetP^{\Theta}\{S\}$ over the set of S instances present in the subset that reached the node to get the average probability on each class:

$$BetP^{\Theta}\{S\}\{C_i\} = \frac{1}{|S|} \sum_{C_i \in C \subseteq \Theta} BetP^{\Theta}\{I_j\}\{C_i\} \quad (3)$$

Compute the entropy of average pignistic probabilities in S:

$$Info(S) = -\sum_{i=1}^{n} BetP^{\Theta}\{S\}\{C_i\} * log_2 BetP^{\Theta}\{S\}\{C_i\} \quad 4$$

where n is a number of possible classes.

For every attribute/feature, the algorithm collects the subset $S_V{}^A$ made with the cases having V as a value for the attribute A, compute pignistic probability $BetP^{\Theta}\{S_V^A\}$ for each v of attribute A. Finally compute $Info_A(S)$ for every attribute as:

$$Info_A(S) = \sum_V \frac{|S_V^A|}{|S|} Info \quad (S_V^A) \quad 5$$

Compute the information gain:

$$Gain(S, A) = Info(S) - Info_A(S) \quad 6$$

and the gain ratio:

$$Gain\ Ratio(S, A) = \frac{Gain(S,A)}{Split\ Info(S,A)} \quad 7$$

Where $Split\ Info(S, A)$ is calculated as:

$$Split\ Info \quad (S, A) = \sum_V \frac{|S_V|}{|S|} * log_2 \frac{|S_V|}{|S|} \quad 8$$

The attribute/feature that produced the largest value of gain ratio is used for the split.

There were several modifications that we introduced into the algorithm.

While the approach described by Elouedi assumes a categorical nature of the attributes, attributes present in LIDC dataset are continuous. We modified the algorithm to work with continuous attributes by setting the threshold on attribute value that will divide a set of instances into the subset. In order to choose an appropriate threshold, we employed the approach proposed by Quinlan in [19]. The approach extracts a separate threshold from every distinct pair of values in the sorted set of attribute values and uses described gain ratio maximization technique to determine the most suitable one.

We also noticed that the Gain Ratio splitting criteria in the case of the LIDC dataset tends to favor very unbalanced splits, assigning a very small ratio of training instances (as small as stopping rules allow) to one of the node's children at every case. As a result the produced trees contained large number of terminal nodes and were over fitted. In order to avoid this we decided to use information gain instead of gain ratio as a splitting criterion.

As the last change we modified one of the stopping rules setting the smallest number of instances that can reach the node to 10 and setting the smallest number of instances that can reach the terminal node to 5. This change has also been done to avoid over fitting of the classification model.

When evaluating a classification system that utilizes a distribution of ratings or classes as an input, and outputs a probability distribution of class membership, evaluation methods beyond accuracy should be used to better capture performance of the system. We propose the idea of a distance curve, in a similar vein to a ROC (receiver operator characteristic) curve, to assess the performance of multiple-label classification approach.

To generate the curve, we varied the thresholds of distance between the distributions for the classification to be considered "accurate." For example, if we looked for nodules that have a normalized distance of 0 between the input and output distributions, we would find little to none. As we increase the distance we find more and more nodules within that threshold. With a normalized distance threshold of 1 between distributions, all the nodules would be considered correct or accurate. Once the curve is generated, the area under the distance threshold curve (AuC_{dt}) was used as the metric for comparison.

An experiment was run to compare the accuracy results of a classification system with the corresponding distance curve AuC_{dt} results. ActiveDECORATE classification [20] on the LIDC dataset was run using 914 instances and the input was the mode of the radiologist ratings as the reference truth. Jeffery Divergence was used as the measure to compare the distance between this input and the classifier output probability distribution. Because the correlation between the accuracy and AuC_{dt}results of the classification system is as high as 0.9947, we can infer that the distance curve AuC_{dt}can be used as an appropriate measure of accuracy to assess the performance of multiple-label classification approach.

4. RESULTS

The dataset used for training and testing the belief decision trees contained 914 instances (1 instance per nodule). The multiple-label of every instance was constructed as class membership probability distribution, where each class probability was calculated as the ratio of radiologists who assigned the nodule to a given class to the total number of radiologists for that nodule. The set of attributes for the instance was generated from the largest (with respect to the area) outline available for a given nodule. The constructed model assigned the predicted multiple-label to each instance at classification step.

The dataset used for training and testing the ActiveDECORATE classification approach contained 2204 instances with instance being a single label assigned to a nodule by a certain radiologist with a set of attributes being generated from the largest outline produced for a nodule by given radiologist. Such a dataset construction approach lead to the fact that each nodule could have up to 4 separate instances associated with it. ActiveDECORATE has also assigned the multiple-label to each instance at the classification step. The experimental design for 3 different classification approaches is summarized in table 4. In order to calculate "nodule based" classification accuracy for the results produced by ActiveDECORATE (to be able to legitimately compare the classification performance of ActiveDECORATE and that of Belief decision trees) the dataset was modified as follows:

The *assigned* instance multiple-label of every instance was once again constructed as a class membership probability distribution, where each class probability was calculated as the ratio of radiologists who assigned the nodule to a given class to the total number of radiologists for that nodule.

The *predicted* multiple-label of every instance was constructed by averaging class membership (class membership → rating) probability distributions produced by ActiveDECORATE for all the instances of given nodule.

Table 4. Experimental design summary for traditional decision trees, ActiveDECORATE and Belief decision trees.

Classification approach	Number of instances	Assigned label	Predicted label.	Vector of attributes
Traditional decision trees	2204 (total number of slices and boundary based images)	A single rating assigned to a nodule by 1 radiologist	Class membership probability distribution for an instance calculated as distribution of instances of each class that reached the particular leaf node of the tree.	Generated from largest outline produced for a nodule by given radiologist
ActiveDECORATE	2204	A single rating assigned to a nodule by 1 radiologist	Class membership probability distribution for an instance calculated by averaging the predictions of ensemble members.	Generated from largest outline produced for a nodule by given radiologist
Belief decision trees	914	A distribution of ratings over panel of radiologists (multiple label)	Class membership probability distribution for an instance calculated by averaging assigned probability labels of the instances that reached the particular leaf node.	Generated from largest outline available for a given nodule.

We divided the dataset into 90% training and 10% testing subsets in such a way that the nodule distributions of testing subsets mimic the nodule distributions of the training subsets with respect to radiologist agreement and the number of radiologists who rated the nodule. Probabilistic trees were grown for each of seven semantic characteristics. Produced trees were then validated on 10% testing subsets. AuC_{dt}values, as well as accuracy values produced by forcing the consensus on assigned and predicted probabilistic labels, for seven semantic characteristics for both 90% and 10% subsets are shown at tables 5 and 6. The plotted curves can be found in Appendix A. Tables 5 and 6 also reports area under the curve and accuracy values for the predicted labels generated by the ActiveDECORATE approach and traditional decision trees. The corresponding columns of both tables report classification accuracies obtained by training and testing the classification model on single label version of the dataset (2204 instances) for traditional decision trees and ActiveDECORATE. Since the belief decision tree algorithm was originally designed for solving multiple-label classification problem it was not possible to obtain classification accuracy values for single label version of the dataset.

Obtained results demonstrate that the belief decision tree approach outperforms both ActiveDECORATE and traditional decision tree algorithms with respect to accuracy and area under the curve (table 7). The average performance boost was 23.88% in comparison with traditional decision trees and 8.19% in comparison with ActiveDECORATE for $AuC_{dt}$and 24.65% and 8.13% for the nodule-based accuracy for the 90% training subset.

Table 5. Comparison of traditional decision tree, ActiveDECPRATE and belief decision tree classification approaches with respect to accuracy (ACC) and area under the distance threshold curve (AuC_{dt}) performance metrics. (90% training subset)

	Traditional Decision Tree			ActiveDECORATE			Belief Decision Tree		
Semantic characteristic	90% subset AuC_{dt} (nodule based)	90% subset ACC (nodule based)	90% subset ACC (2204 instances – previous work)	90% subset AuC_{dt} (nodule based)	90% subset ACC (nodule based)	90% subset ACC (2204 instances – previous work)	90% subset AuC_{dt} (nodule based)	90% subset ACC (nodule based)	90% subset ACC (2204 instances – previous work)
Lobulation	47.46%	41.15%	49.39%	53.90%	43.10%	54.52%	79.97%	69.62%	-/-
Malignancy	42.82%	33.32%	39.44%	87.99%	79.88%	90.65%	73.10%	61.58%	-/-
Margin	37.15%	34.92%	38.54%	66.95%	59.17%	75.62%	70.51%	61.92%	-/-
Sphericity	43.48%	29.76%	33.89%	82.36%	73.02%	86.65%	60.28%	45.93%	-/-
Spiculation	64.60%	50.81%	60.24%	46.93%	31.45%	50.85%	82.05%	74.33%	-/-
Subtlety	42.44%	31.06%	38.87%	77.84%	71.60%	83.35%	70.86%	60.51%	-/-
Texture	73.57%	62.16%	67.26%	45.39%	40.63%	54.32%	81.94%	81.87%	-/-
Average	**50.22%**	**40.46%**	**46.80%**	**65.87%**	**56.98%**	**70.85%**	**74.10%**	**65.11%**	**-/-**

Table 6. Comparison of traditional decision tree, ActiveDecorate and Belief decision tree classification approaches with respect to accuracy (ACC) and area under the distance threshold curve (AUC_DT_dt) performance metrics. (10% testing subset)

	Traditional Decision Tree			ActiveDECORATE			Belief Decision Tree		
Semantic characteristic	10% subset AuC_{dt} (nodule based)	10% subset ACC (nodule based)	10% subset ACC (2204 instances – previous work)	10% subset AuC_{dt} (nodule based)	10% subset ACC (nodule based)	10% subset ACC (2204 instances – previous work)	10% subset AuC_{dt} (nodule based)	10% subset ACC (nodule based)	10% subset ACC (2204 instances – previous work)
Lobulation	30.29%	9.84%	18.60%	50.00%	34.09%	36.41%	74.46%	58.24%	-/-
Malignancy	40.95%	28.81%	31.00%	43.03%	30.85%	35.75%	64.16%	49.45%	-/-
Margin	39.64%	34.51%	36.11%	50.79%	40.66%	46.46%	63.72%	48.91%	-/-
Sphericity	19.61%	6.04%	14.26%	64.90%	56.84%	57.49%	63.14%	37.36%	-/-
Spiculation	34.73%	30.00%	33.53%	35.35%	22.00%	34.92%	76.61%	71.74%	-/-
Subtlety	25.71%	19.46%	25.14%	33.65%	13.40%	15.03%	61.67%	37.36%	-/-

Texture	35.89%	38.53%	40.88%	38.17%	36.59%	47.46%	76.87%	77.17%	-/-
Average	**30.40%**	**23.89%**	**28.50%**	**45.13%**	**33.49%**	**39.07%**	**68.66%**	**54.32%**	**-/-**

Table 7. Summary of performance comparison between ActiveDECORATE and Belief decision trees.

	90%				**10%**			
	AUC_DT Difference		**ACC Difference**		**AUC_DT Difference**		**ACC Difference**	
Semantic characteristic	**BDT vs. DT**	**BDT vs. AD**	**BDT vs. DT**	**BDT vs. AD**	**BDT vs. DT**	**BDT vs. AD**	**BDT vs. DT**	**BDT vs. AD**
Lobulation	32.51%	26.07%	28.47%	26.52%	44.17%	24.46%	48.40%	24.15%
Malignancy	30.28%	-14.89%	28.26%	-18.30%	23.21%	21.13%	20.64%	18.60%
Margin	33.36%	3.56%	27.00%	2.75%	24.08%	12.93%	14.40%	8.25%
Sphericity	16.80%	-22.08%	16.17%	-27.09%	43.53%	-1.76%	31.32%	-19.48%
Spiculation	17.45%	35.12%	23.52%	42.88%	41.88%	41.26%	41.74%	49.74%
Subtlety	28.42%	-6.98%	29.45%	-11.09%	35.96%	28.02%	17.90%	23.96%
Texture	8.37%	36.55%	19.71%	41.24%	40.98%	38.70%	38.64%	40.58%
Average	**23.88%**	**8.19%**	**24.65%**	**8.13%**	**36.26%**	**23.53%**	**30.43%**	**20.83%**

Finally we noticed that for 10% training subset belief decision trees outperform ActiveDECORATE for 6 semantic characteristics out of 7 which suggest high generalization of the belief decision tree classification approach.

From table 5 we notice that belief decision trees outperforms ActiveDECORATE (Lobulation – 26.52% increase, Spiculation – 42.88% increase, Texture – 41.24% increase) on those semantic characteristics for which a highly dominant rating exists (Appendix B).

In order to determine the impact of a distribution's shape on classification accuracy the subsets of correctly classified and misclassified instances were examined independently (Appendix C). For semantic characteristics mentioned in the previous paragraph we determined that belief decision trees accurately predicted the majority of instances with dominant rating while the performance of ActiveDECORATE for these instances was always <50% for all 3 semantic characteristics. However, the situation was the opposite for instances with non-dominant ratings, on which ActiveDECORATE has always performed better than belief decision trees. The ActiveDECORATE vs. belief decision trees increase in performance for those instances was 6.6% for lobulation, 4.4% for speculation and 0.4% for texture. A summary of our findings is reported in table 8.

Table 8. Misclassification rate of ActiveDecorate and Belief decision tree classification approaches on instances with dominant and non-dominant ratings. (90% training subset)

		ActiveDECORATE				**Belief Decision Tree**			
	Characteristic	**Number of correctly classified DOMINANT instances**	**Number of misclassified DOMINANT instances**	**Number of correctly classified NON-DOMINANT instances**	**Number of misclassified NON-DOMINANT instances**	**Number of correctly classified DOMINANT instances**	**Number of misclassified DOMINANT instances**	**Number of correctly classified NON-DOMINANT instances**	**Number of misclassified NON-DOMINANT instances**

Counts	Lobulation	252	359	101	111	553	58	87	125
	Spiculation	212	451	44	115	627	36	37	122
	Texture	261	314	77	170	565	10	76	171
%	Lobulation	30.62%	43.62%	12.27%	13.49%	67.19%	7.05%	10.57%	15.19%
	Spiculation	25.79%	54.87%	5.35%	13.99%	76.28%	4.38%	4.50%	14.84%
	Texture	31.75%	38.20%	9.37%	20.68%	68.73%	1.22%	9.25%	20.80%

In order to understand the reasons for such behavior of these two approaches it is necessary to clearly determine how the instance label affects the process of building the classification model.

For belief decision trees the instance label is used to calculate the average pignistic probability function (average across 5 classes) which is then used for calculating the entropy of the set and determining the goodness of split for a particular node. Every node in a belief decision tree has a probability distribution associated with it which is calculated by averaging theprobability distributions (uncertain labels) of instances that reach that node during the training phase. At the classification step, a classified instance is assigned the probability distribution of a leaf node that it reaches. It is obvious that since all predicted labels are produced by averaging the subset of assigned instance labels and there exists a rating which is highly dominant across all 5, there will be fair amount of predicted uncertain labels with the given rating also being dominant. Due to the way accuracy is assessed for every case (mode vs. mode) the model will perform well for instances with dominant ratings.

For ActiveDECORATE algorithm the uncertain label for instance being classified is calculated using equation 9

$$P_{y_k}(x) = \frac{\sum_{\forall_i (C_i \in C*)} P_{C_i, y_k}(x)}{\sum_{\forall_i (C_i \in C*), \forall_j (y_j \in Y)} P_{C_i, y_j}(x)} \qquad 9$$

Where $P_{y_k}(x)$ is probability for class y_k and $P_{C_i,y_k}(x)$ is probability of instance to be of class y_k according to a particular classifier Ci of the ensemble.

The uncertain label of an instance (instance in this case being a single radiologist's diagnosis for a given nodule) therefore calculated as follows:

Each of the 5 (in case of LIDC dataset) class probabilities is calculated by summing up probabilities produced for this class by all classifiers in the ensemble and dividing the result by the number of classifiers in the ensemble. After probability distribution is calculated for every instance of a nodule, probability distributions of these instances are averaged together to produce the predicted label of a nodule. The way in which a probability distribution is initially generated by a single base classifier depends on a nature of the base classifier. In particular C4.5 decision tree calculates class probabilities of a label on a leaf node from a ratio of instances of every particular class that reached that node.

Consider an example (table 9) in which there are 3 instances that reached a leaf node of a decision tree at classification step. Depending on the type of a decision tree (traditional vs. belief) the multiple-label of the node will be calculated differently. While traditional decision tree will take only mode ratings into

consideration thus ignoring 1 indeterminate rating assigned to nodule 3, belief decision tree will take all the ratings assigned to all 3 nodules into account.

Table 9. Example of calculating node's label for traditional and belief decision tree.

MALIGNANCY	Highly Unlikely	Moderately Unlikely	Indeterminate	Moderately Suspicious	Highly Suspicious
Nodule 1	0.25	0.75	0	0	0
Nodule 2	0.75	0.25	0	0	0
Nodule 3	0.75	0	0.25	0	0
ActiveDECORATE	0.66	0.34	0	0	0
Belief decision tree	0.58	0.33	0.08	0	0

5. CONCLUSION

We were able to adopt, apply and evaluate multiple-label classification algorithm for classifying lung nodules contained in LIDC dataset. To the best of our knowledge, in all previous multiple-label classification research works authors worked with synthetic dataset created or collected for the purpose of testing the described technique. We were able to achieve average classification accuracy of 69% across seven semantic characteristics using the area under the curve evaluation criteria. Our future work will extend in two directions: first we will employ both 3D image information and information from multiple radiologists outlines to generate a set of image features to improve classification performance, second we will attempt to adopt different classification techniques such as ANN or SVM to multiple-label classification problems to determine the best in terms of classification performance and incorporate optimal technique into the active learning [21] approach described in our previous work.

REFERENCES

1 Dave P. Miller, Kathryn F. O'Shaughnessy, Susan A.Wood, Ronald A. Castellino, "Gold Standards and Expert Panels: A Pulmonary Nodule Case Study with Challenges and Solutions" SPIE Medical Imaging, 2004

2 Hu, Z. Li, Y. Cai, Y. Xu, X., Method of Combining Multi-Class SVMs Using Dempster-Shafer Theory and Its Application, PROCEEDINGS OF THE AMERICAN CONTROL CONFERENCE, 2005, VOL 3, pages 1946-1950

3 T. G. Dietterich, R. H. Lathrop, and T. L.-Perez (1997) Solving the multiple-instance problem with axis-parallel rectangles, Artificial Intelligence, 89(1-2), pp. 31-71.

4 Li, T., Zhang, C., and Zhu, S. 2006. “Empirical Studies on Multi-label Classification.” In Proceedings of the 18th IEEE international Conference on Tools with Artificial intelligence (November 13 - 15, 2006). ICTAI. IEEE Computer Society, Washington

5 Ghamrawi, N. and McCallum, A. 2005. Collective multi-label classification. In *Proceedings of the 14th ACM international Conference on information and Knowledge Management* (Bremen, Germany, October 31 - November 05, 2005). CIKM '05. ACM, New York, NY, 195-200.

6 Ramachandran, C., Malik, R., Jin, X., Gao, J., Nahrstedt, K., and Han, J. “VideoMule: a consensus learning approach to multi-label classification from noisy user-generated videos.” In Proceedings of the Seventeen ACM international Conference on Multimedia (Beijing, China, October 19 - 24, 2009). MM '09. ACM, New York, NY, 721-724.

7 Xipeng Shen, Matthew Boutell, Jiebo Luo, and Christopher Brown, ``Multi-label Machine Learning and Its Application to Semantic Scene Classification'', In Proceedings of IS&T/SPIE's Sixteenth Annual Symposium on Electronic Imaging: Science and Technology (EI 2004), San Jose, California, USA, January 2004, 188—199

8 Cheng, W. and Hüllermeier, E. “Combining instance-based learning and logistic regression for multi-label classification.” Machine Learning. 76, 2-3 (Sept. 2009), 211-225.

9 A. Veloso , W. Meira, Jr. , M. Gonçalves , M. Zaki, Multi-label Lazy Associative Classification, Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases, September 17-21, 2007, Warsaw, Poland [doi>10.1007/978-3-540-74976-9_64]

10 Jin, R. & Ghahramani, Z. 2002, Learning with Multiple Labels, paper presented to Proceedings of Neural Information Processing Systems 2002 (NIPS 2002), Vancoucer, Canada.

11 Tsoumakas, G., Katakis, I. 2007, Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3, 1–13

12 M.S. Bjanger and T. Denœux, Induction of decision trees from partially classified data using belief functions, Master, University of Compiegne, 2000.

13 T. Denoeux, A Neural Network Classifier Based on Dempster-Shafer Theory, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 30, NO. 2, MARCH 2000

14 Quost, B. and Denœux, T. 2009. Learning from data with uncertain labels by boosting credal classifiers. In *Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery From Uncertain Data* (Paris, France, June 28 - 28, 2009). J. Pei, L. Getoor, and A. de Keijzer, Eds. U '09. ACM, New York, NY, 38-47. DOI= http://doi.acm.org/10.1145/1610555.1610561

15 G. Shafer. A mathematical theory of evidence. Princeton University Press, Princeton, NJ, 1976.

16 P. Vannoorenberghe, T. Denoeux, “Handling Uncertain Labels in Multiclass Problems using Belief Decision Trees”, International Conference on Processing and Management of Uncertainty, pp. 1916-1926, 2002, Annecy, France, 2002.

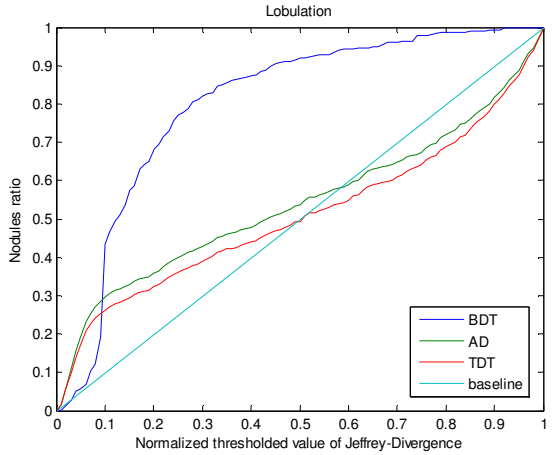
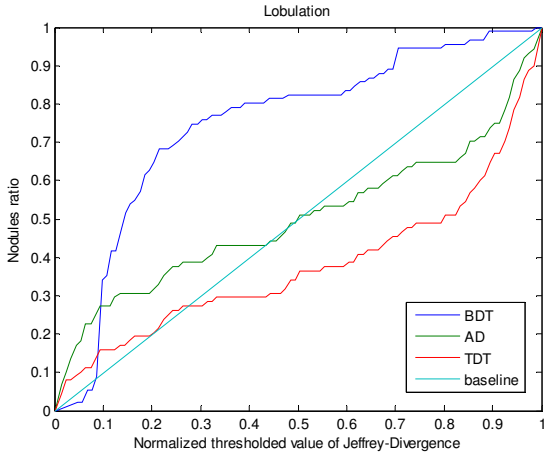
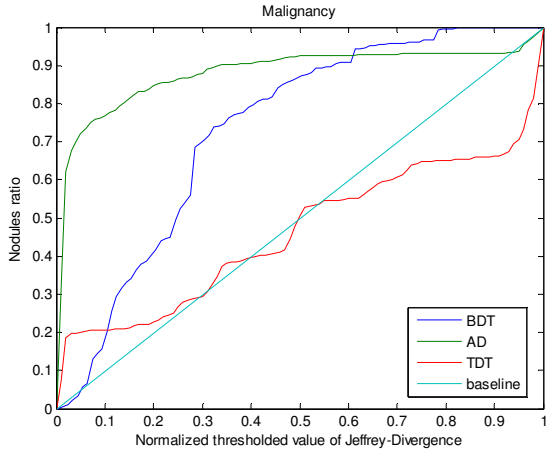
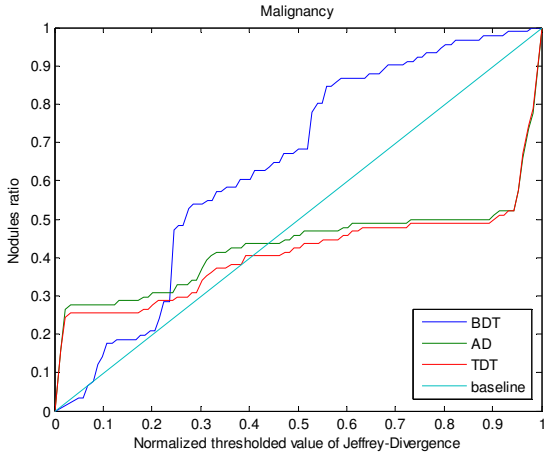
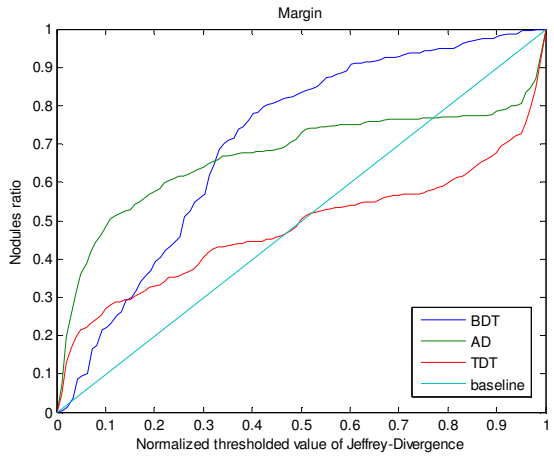
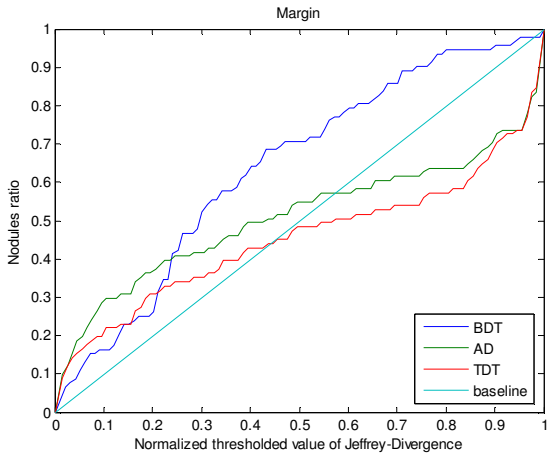
17 G. Armato III S. et al, Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community, doi: 10.1148/radiol.2323032035, September 2004 Radiology, 232, 739-748.

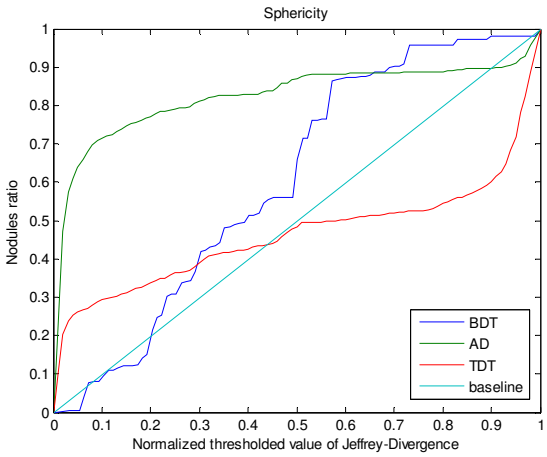
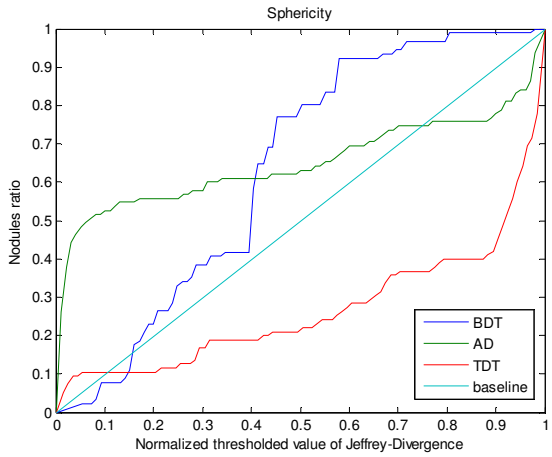
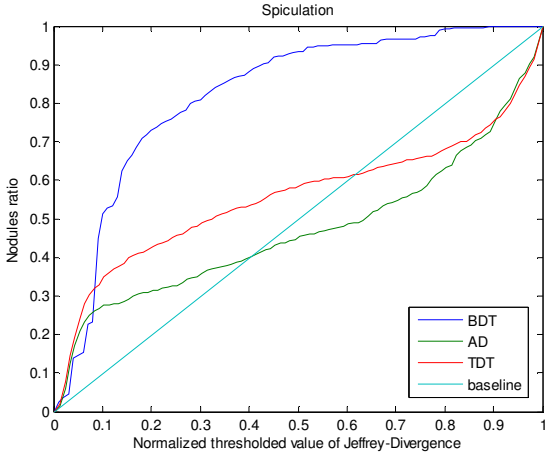
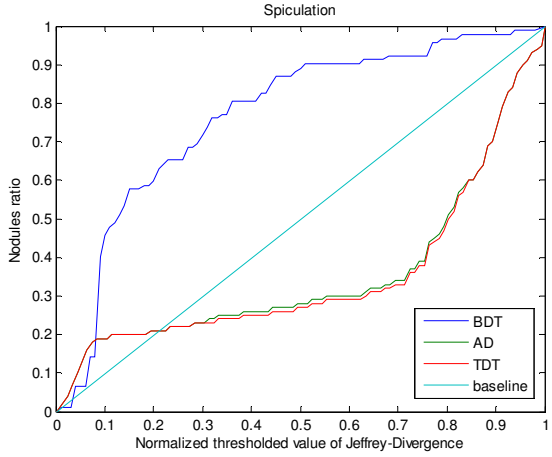
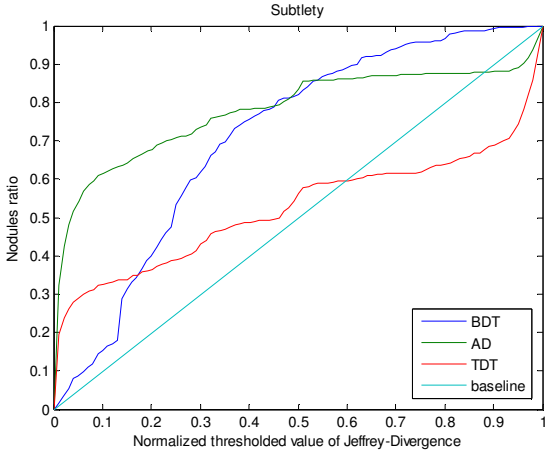
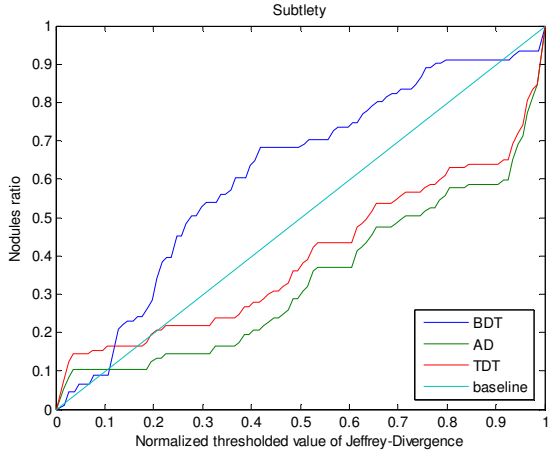
18 Z. Elouedi, K. Mellouli, and P. Smets. Belief decision trees: Theoretical foundations. International Journal of Approximate Reasoning, 28:91–124, 2001.

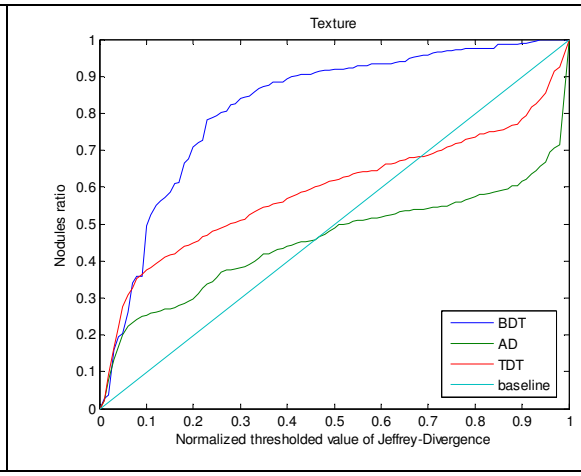
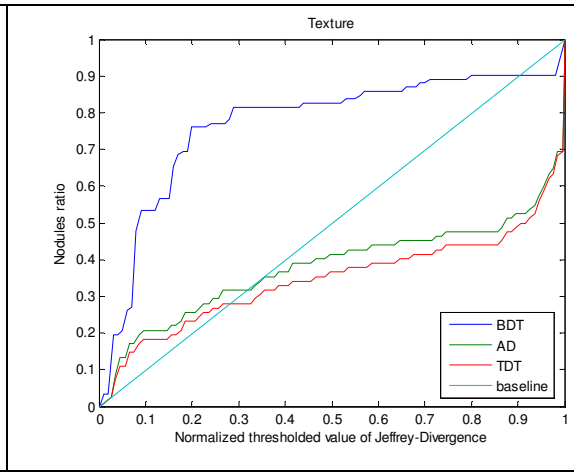
19 J. R. Quinlan. Improved Use of Continuous Attributes in C4.5. Journal of Articial Intelligence Research 4 (1996) 77-90

20 Zinovev, D.; Raicu, D.; Furst, J.; Armato III, S.G. Predicting Radiological Panel Opinions Using a Panel of Machine Learning Classifiers. Algorithms 2009, 2, 1473-1502.
21 Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.

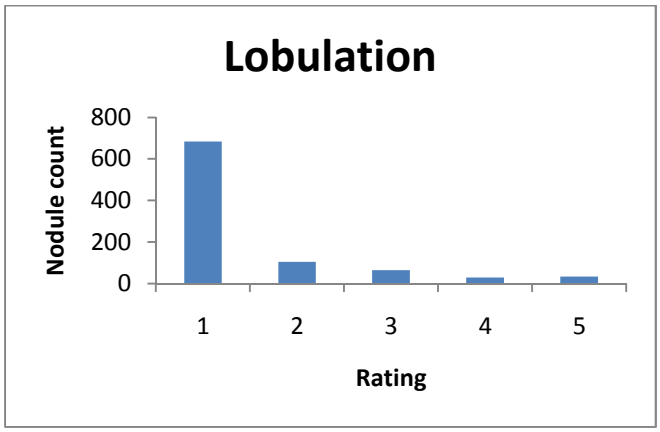
APPENDIX A. Jeffrey divergence curves produced using belief decision trees

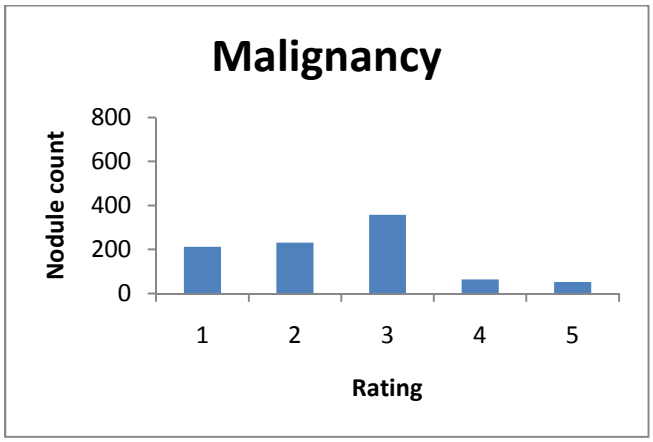
Semantic characteristic	90% subset AUC	10% subset AUC
Lobulation		
Malignancy		
Margin		

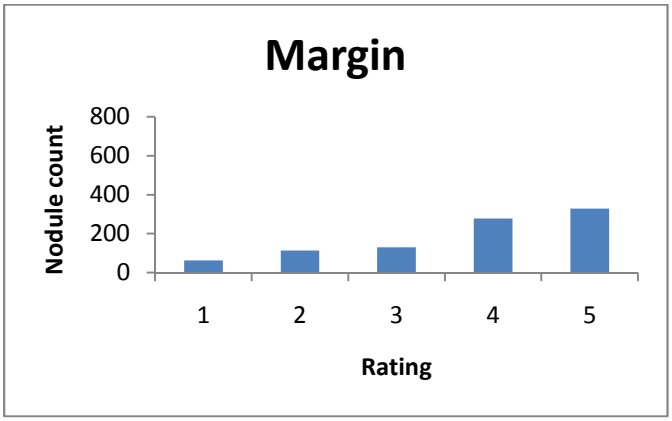
Sphericity		
Spiculation		
Subtlety		

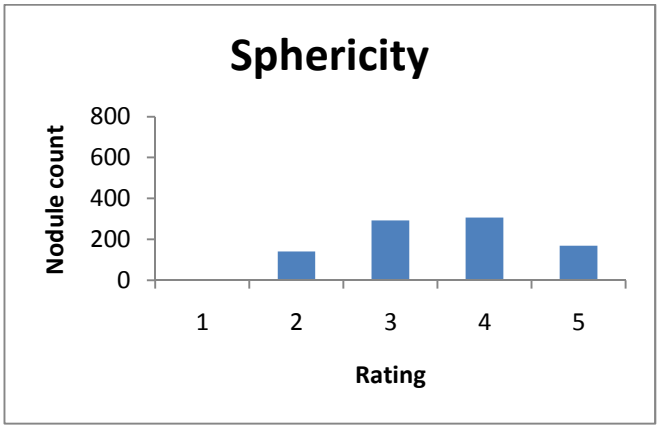
Texture		

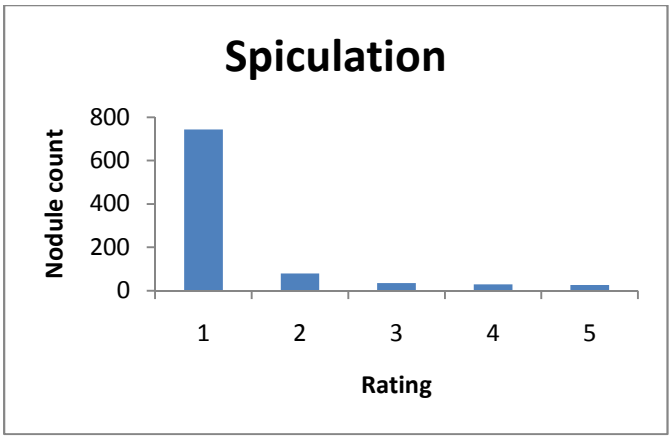
APPENDIX B. Distribution of ratings in 90% training subset for 7 semantic characteristics (Rating of a nodule is defined as mode of radiologist ratings assosiated with a nodule)

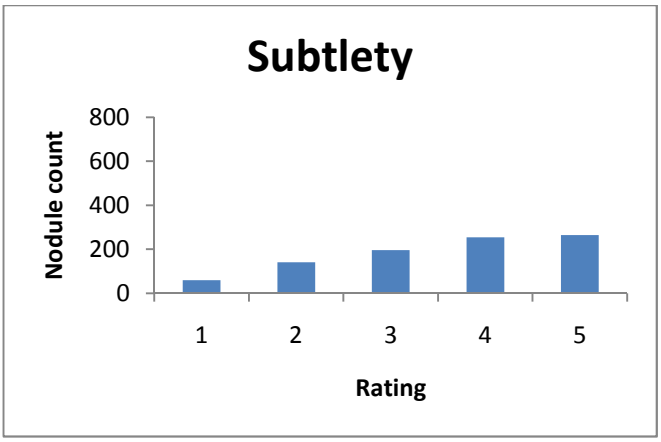


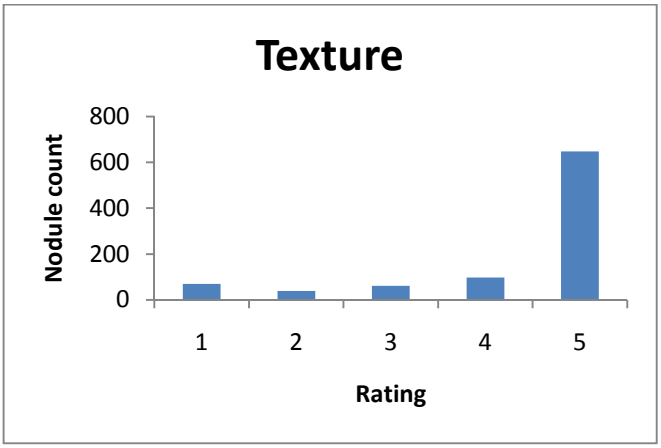




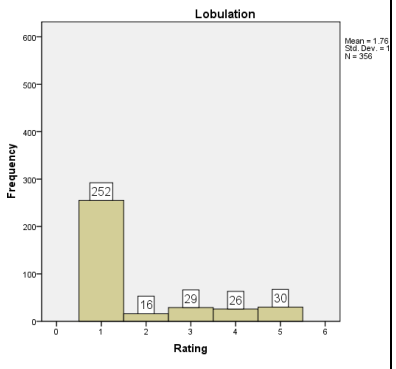
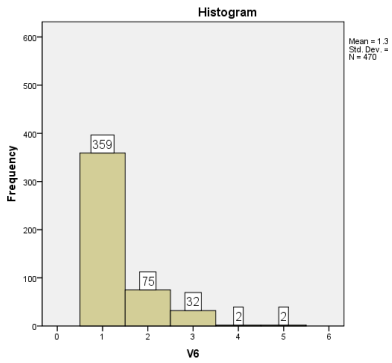
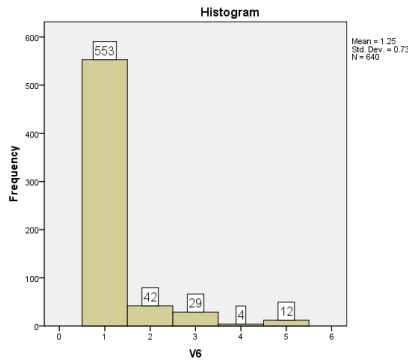
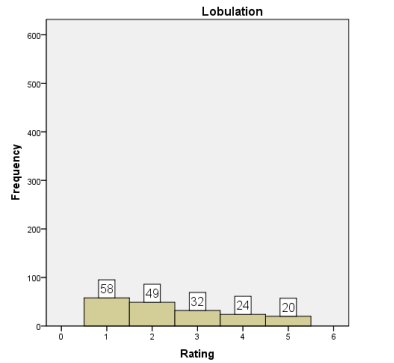
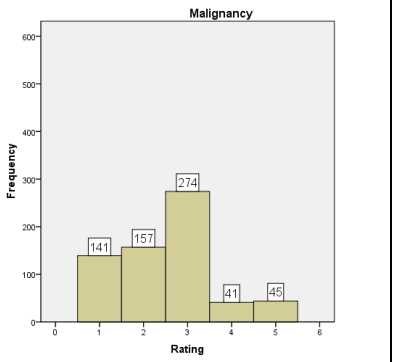
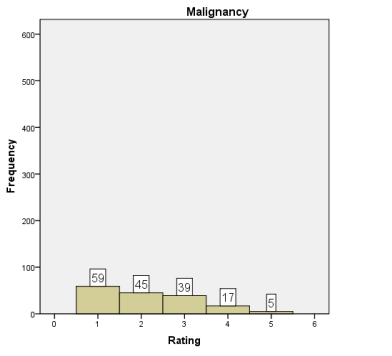
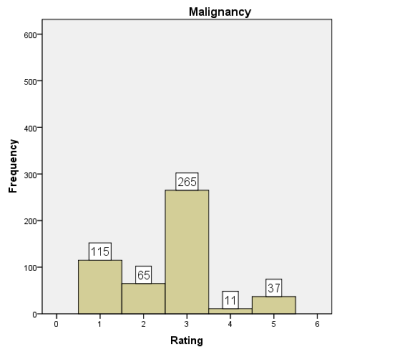
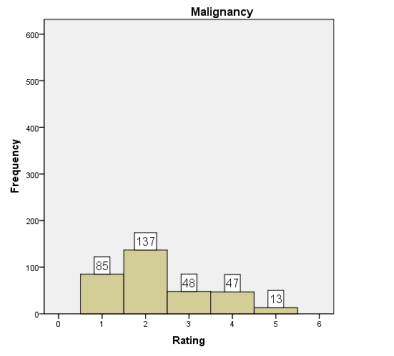
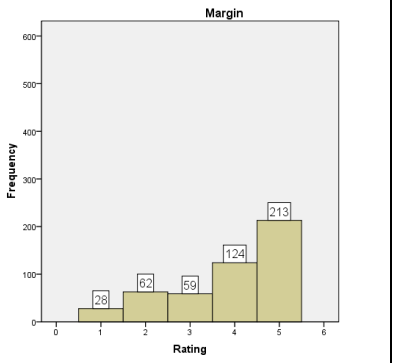
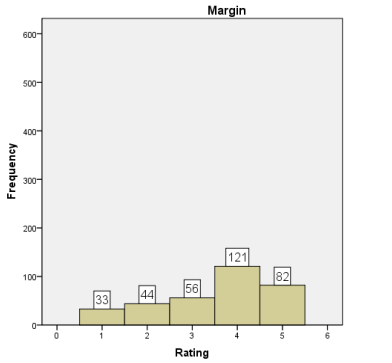
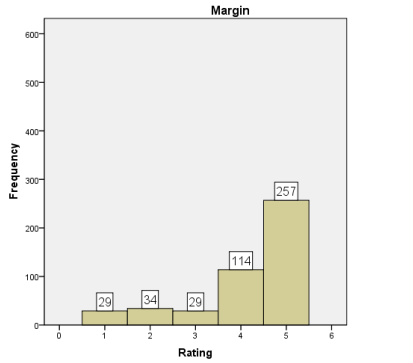
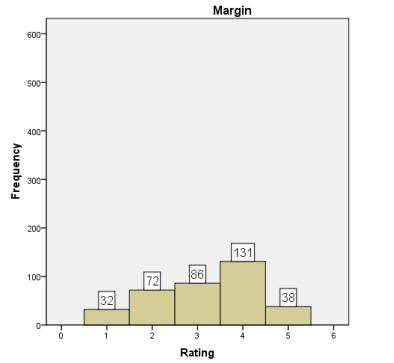


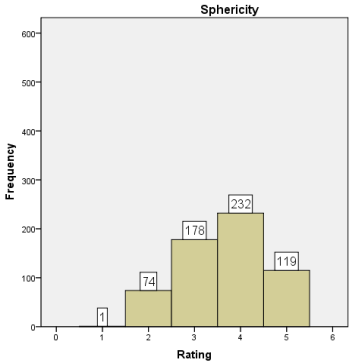
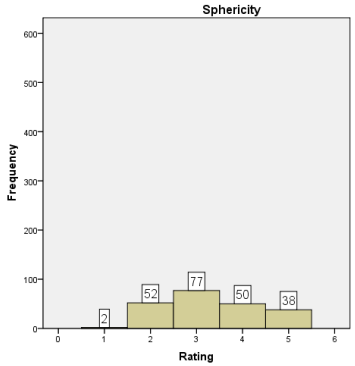
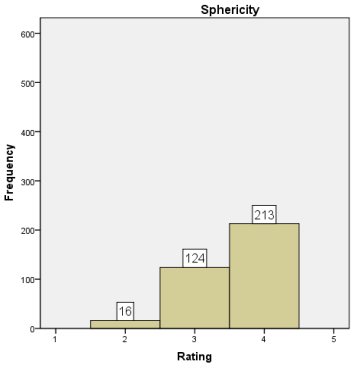
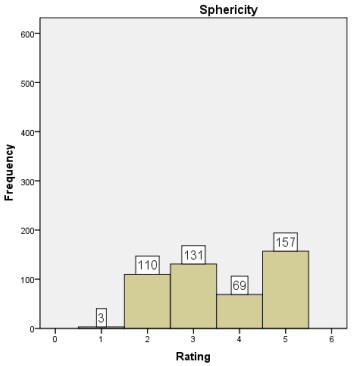
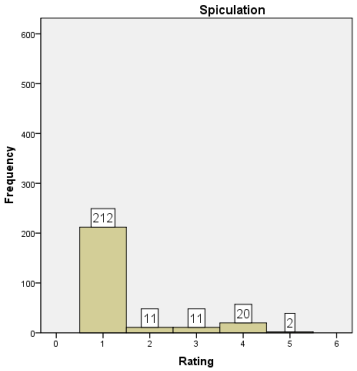
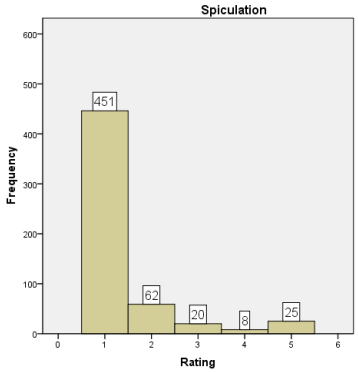
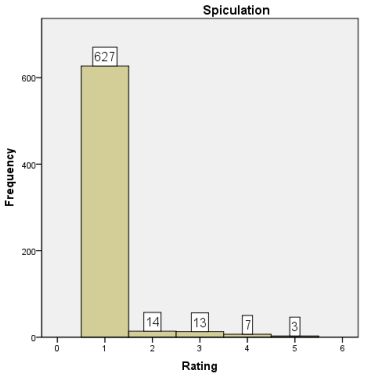
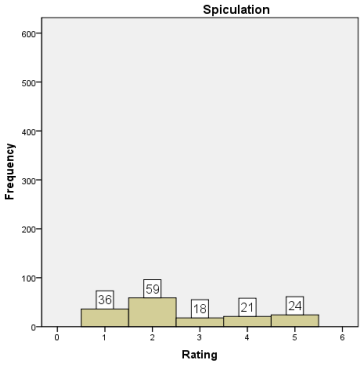
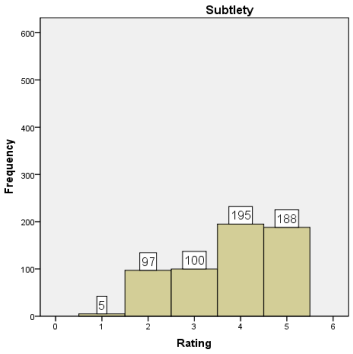
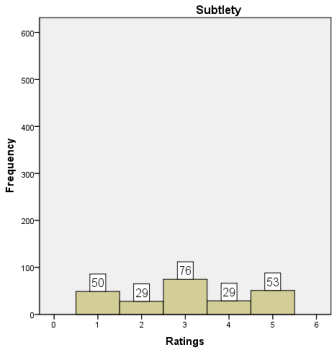
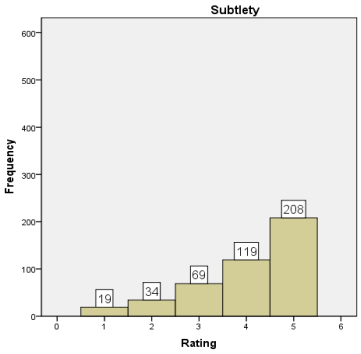
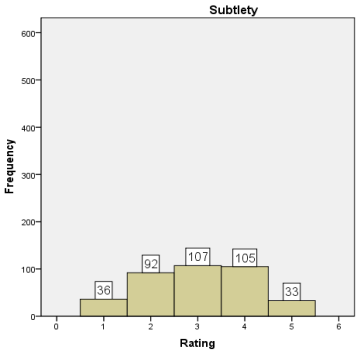

Spiculation
Nodule count
800
600
400
200
0
1
2
3
4
5
Rating

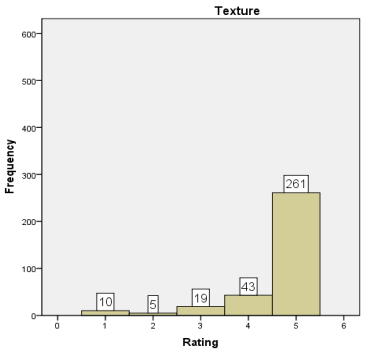
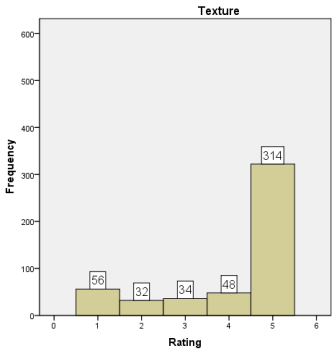
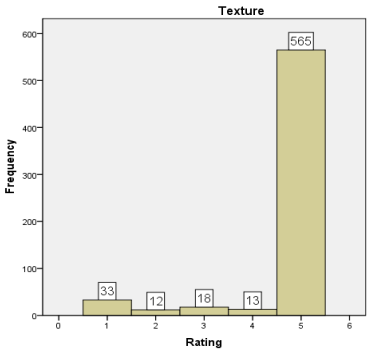
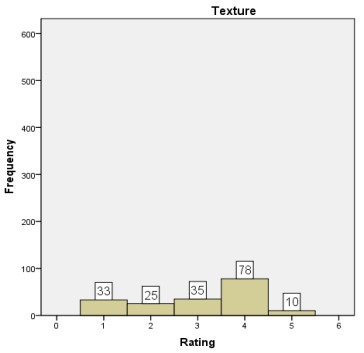

Subtlety
Nodule count
800
600
400
200
0
1
2
3
4
5
Rating


Texture
Nodule count
800
600
400
200
0
1
2
3
4
5
Rating

APPENDIX C. Distribution of ratings in subsets of correctly and incorrectly classified instances from 90% training set for ActiveDECORATE algorithm and Belief decision trees. (Rating of a nodule is defined as mode of radiologist ratings assosiated with a nodule)

	AD		BDT	
	Correctly classified	Misclassified	Correctly classified	Misclassified
Lobulation				
Malignancy				
Margin				

Sphericity				
Spiculation				
Subtlety				

Texture	 Texture Frequency Rating 10 5 19 43 261	 Texture Frequency Rating 56 32 34 48 314	 Texture Frequency Rating 33 12 18 13 565	 Texture Frequency Rating 33 25 35 78 10