International Journal of Advanced Engineering Application
Volume No.2 Issue No 1 Jan 2025
ISSN NO:3048-6807

Forecasting Radiological Panel Opinions Through Ensemble Machine Learning Classifiers

João M. Santos[1], Ana P. Silva[2], and Carlos F. Mendes[3]
[1]Department of Electrical Engineering, Agostinho Neto University, Luanda, Angola
[2]Department of Computer Science, Agostinho Neto University, Luanda, Angola
[3]Department of Civil Engineering, Agostinho Neto University, Luanda, Angola

Abstract:
This paper explores the use of an ensemble of machine learning classifiers combined with active learning strategies to predict radiologists' assessments of lung nodule characteristics in the Lung Image Database Consortium (LIDC). The study focuses on modeling and predicting agreement among radiologists' semantic ratings across seven key nodule characteristics: spiculation, lobulation, texture, sphericity, margin, subtlety, and malignancy. These characteristics are essential in evaluating and diagnosing pulmonary nodules.The proposed approach utilizes an ensemble of classifiers, functioning as a simulated "computer panel of experts," to analyze 64 image features extracted from the nodules. These features span four critical categories: shape, intensity, texture, and size. By leveraging active learning, the system initiates the training phase with nodules where radiologists' semantic ratings are consistent. The system then progressively learns to classify nodules with varying degrees of disagreement among radiologists, effectively addressing uncertainty and variability in expert interpretations. The results demonstrate that the ensemble approach outperforms individual classifiers in terms of classification accuracy, showcasing its ability to synthesize diverse perspectives and make more reliable predictions. This enhanced predictive capability underscores the potential of machine learning to serve as a supportive tool in radiological diagnostics. By acting as a "second read" for physicians, the proposed system can improve consistency in radiological interpretations, reduce diagnostic variability, and ultimately enhance patient care. The findings highlight the promising role of advanced computational methods in augmenting human expertise in medical imaging analysis.

Keywords: Ensemble learning; Active learning; Lung nodule classification; LIDC database; Radiological assessment; Semantic characteristics; Nodule spiculation; Nodule lobulation; Nodule texture; Nodule sphericity; Nodule margin; Nodule subtlety; Malignancy prediction; Machine learning in radiology; Image feature analysis

1. Introduction

Interpretation performance varies greatly among radiologists when assessing lung nodules on computed tomography (CT) scans. A good example of such variability is the Lung Image Database Consortium (LIDC) dataset [**1**] for which out of 914 distinct nodules identified, delineated, and semantically characterized by up to four different radiologists, there are only 180 nodules on average across seven semantic characteristics on which at least three radiologists agreed with respect to the semantic label (characteristic rating) applied to the nodule. Computer-aided diagnosis (CADx) systems can act as a second reader by assisting radiologists in interpreting nodule characteristics in order to improve their efficiency and accuracy. In our previous work [**2**] we developed a semi-automatic active-learning approach [**3**] for predicting seven lung nodule semantic characteristics: spiculation, lobulation, texture, sphericity, margin, subtlety, and malignancy. The approach was intended to handle the large variability among interpretations of the same nodule by different radiologists. Using nodules with a high level of agreement as initial training data, the algorithm automatically labeled and added to the training data those nodules which had inconsistency in their interpretations. The evaluation of the algorithm was performed on the LIDC dataset publicly available at the time of publication, specifically on 149 distinct nodules present in the CT scans of 60 patients. A new LIDC dataset consisting of 914 distinct nodules from 207 patients was made publicly available as of June 2009. This has opened the way to further investigate the robustness of our proposed approach. Given the highly non-normal nature of medical data in general and of the LIDC dataset in particular (for example, on the set of 236 nodules for which at least three radiologists agree with respect to the spiculation characteristic, 231 of these nodules are rated with a 1 (”marked spiculation”) and only five nodules are rated with ratings from 2 to 5 (where 5 “no spiculation”), we include in our research design a new study to evaluate the effects of balanced and unbalanced datasets on the proposed ensemble’s performance for each of the seven characteristics. Furthermore, we investigate the agreement between our proposed computer-aided diagnostic characterization (CADc) approach and the LIDC radiologists’ semantic characterizations using the weighted kappa statistic [**4**] which takes into account the general magnitude of the radiologists’ agreement and weighs the differences in their disagreements with respect to every available

instance. Finally, we include a new research study to investigate the effects of the variation/disagreement present in the manual lung nodule delineation/segmentation on performance of the ensemble of classifiers.

The rest of the paper is organized as follows: we present a literature review relevant to our work in **Section 2**, the National Cancer Institute (NCI) LIDC dataset and methodology in **Section 3**, the results in **Section 4**, and our conclusions and future work in **Section 5**.

2. Related Work

A number of CAD systems have been developed in recent years for automatic classification of lung nodules. McNitt-Gray *et al*. [**5**,**6**] used nodule size, shape and co-occurrence texture features as nodule characteristics to design a linear discriminant analysis (LDA) classification system for malignant versus benign nodules. Lo *et al*. [**7**] used direction of vascularity, shape, and internal structure to build an artificial neural network (ANN) classification system for the prediction of the malignancy of nodules. Armato *et al*. [**8**] used nodule appearance and shape to build an LDA classification system to classify pulmonary nodules into malignant versus benign classes. Takashima *et al*. [**9**,**10**] used shape information to characterize malignant versus benign lesions in the lung. Shah *et al*. [**11**] compared the malignant *vs.* benign classification performance of OneR [**12**] and logistic regression classifiers learned on 19 attenuation, size, and shape image features; Samuel *et al*. [**13**] developed a system for lung nodule diagnosis using Fuzzy Logic. Furthermore, Sluimer *et al*. [**14**] and more recently Goldin *et al*. [**15**] summarized in their survey papers the existing lung nodule segmentation and classification techniques.

There are also research studies that use clinical information in addition to image features to classify lung nodules. Gurney *et al.* [**16**,**17**] designed a Bayesian classification system based on clinical information, such as age, gender, smoking status of the patient, *etc*., in addition to radiological information. Matsuki *et al.* [**18**] also used both clinical information and sixteen features scored by radiologists to design an ANN for malignant versus benign classification. Aoyama *et al.* [**19**] used two clinical features in addition to forty-one image features to determine the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. Although the work cited above provides convincing evidence that a combination of image features can indirectly encode radiologists' knowledge about indicators of malignancy (Sluimer *et al.* [**14**]), the precise mechanism by which this correspondence happens is unknown. To understand this mechanism, there is a need to explore several approaches for finding the relationships between the image features and radiologists' annotations. Kahn *et al.* [**20**] emphasized recently the importance of this type of research; the knowledge gathered from the post-processed images and its incorporation into the diagnosis process could simplify and accelerate the radiology interpretation process. Notable work in this direction is the work by Barb *et al*. [**21**] and Ebadollahi *et al*. [**22**,**23**]. Barb *et al.* proposed a framework that uses semantic methods to describe visual abnormalities and exchange knowledge in the medical domain. Ebadollahi *et al.* proposed a system to link the visual elements of the content of an echocardiogram (including the spatial-temporal structure) to external information such as text snippets extracted from diagnostic reports. Recently, Ebadollahi *et al.* demonstrated the effectiveness of using a semantic concept space in multimodal medical image retrieval. In the CAD domain, there is some preliminary work to link images to BI-RADS. Nie *et al.* [**24**] reported results linking the gray-level co-occurrence matrix (GLCM) entropy and GLCM sum average to internal enhancement patterns (homogenous versus heterogeneous) defined in BI-RADS, while Liney *et al.* [**25**] linked complexity and convexity image features to the concept of margin and circularity to the concept of shape. Our own work [**26**,**27**] can also be considered one of the initial steps in the direction of mapping lung nodule image features first to perceptual categories encoding the radiologists' knowledge about lung interpretation and further to the RadLex lexicon [**28**].

In this paper we propose a semi-supervised probabilistic learning approach to deal with both the inter-observer variability and the small set of labeled data (annotated lung nodules). Given the ultimate use of our proposed approach as a second reader in the radiology interpretation process, we investigate the agreement between the ensemble of classifiers and the LIDC panel of experts as well as the performance accuracy of the ensemble of classifiers. The accuracy of the ensemble is calculated as the number of correctly classified instances over the total number of instances. The agreement is measured using weighted kappa statistic as introduced by Cohen [**4**,**29**]. The weighted kappa statistic takes into account the level of disagreement and the specific category on which raters agreed for each observed case, reflecting the importance of a certain rating. Originally, the kappa statistic was intended to measure the agreement between two raters across a number of cases, where the pair of raters is fixed for all cases. Fleiss [**30**] proposed a generalization of kappa statistics which measures the overall agreement across multiple observations when more than two raters were interpreting a specific case. Landis and Koch [**31**] explored the use of kappa statistics for assessing the majority agreement by modifying the unified agreement evaluation approach that they proposed in a previously published paper [**32**]. An approach proposed by Kraemer [**33**] extended the technique proposed by Fleiss [**34**] to situations in which there are a multiple number of

observations per subject and a multiple, inconstant number of possible responses per observation. More recently, Viera and Garrett [**35**] published a paper that describes and justifies a possible interpretation scale for the value of kappa statistics obtained in the evaluation of inter-observer agreement. They propose to split the range of possible values of the kappa statistic into several intervals and assign an ordinal value to each of them as shown in **Table 1**. We will use this interpretation scale to quantify the agreement between the panel of LIDC experts and the ensemble of classifiers.

3. Methodology

3.1. LIDC dataset

The publicly available LIDC database (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 scanner models resulting in a wide range of values of the imaging acquisition parameters. For example, slice thickness ranges between 0.6 mm and 4.0 mm, reconstruction diameter ranges between 260 mm and 438 mm, exposure ranges between 3 ms and 6,329 ms, and the reconstruction kernel has one of the following values: B, B30f, B30s, B31f, B31s, B45f, BONE, C, D, FC01, or STANDARD.

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 up to 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 (**Table 2**): subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, and malignancy likelihood. For example, the texture characteristic provides meaningful information regarding nodule appearance ("Non-Solid", "Part Solid/(Mixed)", "Solid") while malignancy characteristic captures the likelihood of malignancy ("Highly Unlikely", "Moderately Unlikely", "Indeterminate", "Moderately Suspicious", "Highly Suspicious") as perceived by the LIDC radiologists. The process by which the LIDC radiologists reviewed CT scans, identified lesions, and provided outlines and characteristic ratings for nodules > 3 mm has been described in detail by McNitt-Gray *et al.* [**36**].

The nodule outlines and the seven of the nodule characteristics were used extensively throughout this study. Note that the LIDC did not impose a forced consensus; rather, all of the lesions indicated by the radiologists at the conclusion of the unblinded reading sessions were recorded and are available to users of the database. Accordingly, each lesion in the database considered to be a nodule > 3 mm could have been marked as such by only a single radiologist, by two radiologists, by three radiologists, or by all four LIDC radiologists. For any given nodule, the number of distinct outlines and the number of sets of nodule characteristic ratings provided in the XML files would then be equal to the number of radiologists who identified the nodule.

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 64 two-dimensional (2D), low-level image features grouped into four categories: shape features, texture features, intensity features, and size features (**Table 3** and **Appendix 1**). Although each nodule is present in a sequence of slices, in this paper 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 (**Figure 1**). In our future work, we will also investigate the use of three-dimensional (3D) features to encode the image content of the lung nodules and compare the classification power of the 3D features versus the 2D features [**37**].

After completion of the feature extraction process, we created a vector representation of every nodule image which consisted of 64 image features and 9 radiologists' annotations (**Figure 2**).

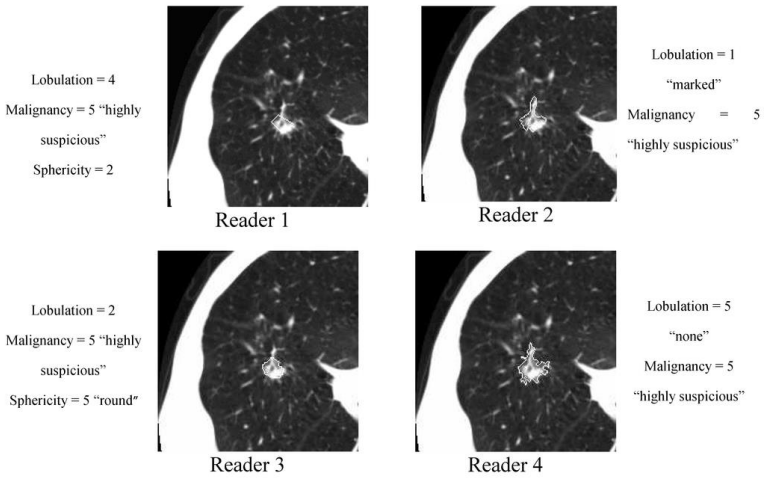

Figure 1. An example of four different delineations of a nodule on a slice marked by four different radiologists.

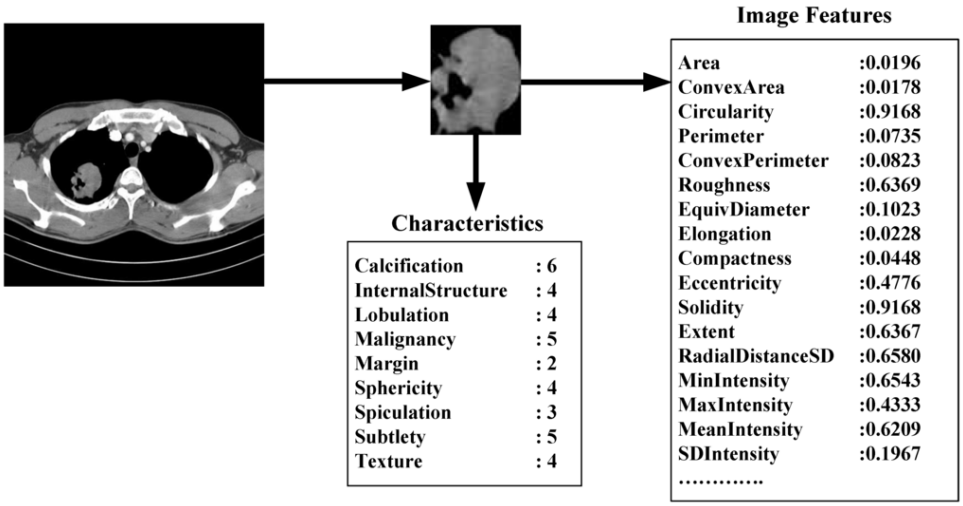

Figure 2. An example of nodule characteristics assigned by a radiologist and normalized low-level features computed from image pixels.

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 eight common image shape features: circularity, roughness, elongation, compactness, eccentricity, solidity, 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).

Co-occurrence matrices focus on the distributions and relationships of the gray-level intensity of pixels in the image. They are calculated along four directions (0°, 45°, 90°, and 135°) and five distances (1, 2, 3, 4 and 5 pixels) producing 20 co-occurrence matrices. Once the co-occurrence matrices are calculated, eleven Haralick texture descriptors are then calculated from each co-occurrence matrix. Although each Haralick texture descriptor is calculated from each co-occurrence matrix, we averaged the features across all distance/direction pairs resulting in 11 (instead of 11 × 4 × 5) Haralick features per image.

Gabor filtering is a transform based method which extracts texture information from an image in the form of a response image. A Gabor filter is a sinusoid function modulated by a Gaussian and discretized over orientation and frequency. We convolve the image with 12 Gabor filters: four orientations (0°, 45°, 90°, and 135°) and three frequencies (0.3, 0.4, and 0.5), where frequency is the inverse of wavelength. We then calculate means and standard deviations from the 12 response images resulting in 24 Gabor features per image.

Markov Random Fields (MRFs) is a model based method which captures the local contextual information of an image. We calculate five features corresponding to four orientations (0°, 45°, 90°, 135°) along with the variance. We calculate feature vectors for each pixel by using a 9 estimation window. The mean of four different response images and the variance response image are used as our five MRF features.

3.3. Active DECORATE for lung nodule interpretation

We propose to find mappings based on a small labeled initial dataset that, instead of predicting a certain rating (class) for a semantic characteristic, will generate probabilities for all possible ratings of that characteristic. Our proposed approach is based on the DECORATE [**38**] algorithm, which iteratively constructs an ensemble of classifiers by adding a small amount of data, artificially generated and labeled by the algorithm, to the data set and learning a new classifier on the modified data. The newly created classifier is kept in the ensemble if it does not decrease the ensemble's classification accuracy. Active-DECORATE [**39**] is an extension of the DECORATE algorithm that detects examples from the unlabeled pool of data that create the most disagreement in the constructed ensemble and adds them to the data after manual labeling. The procedure is repeated until a desired size of the data set or a predetermined number of iterations is reached. The difference between Active-DECORATE and our approach lies in the way examples from the unlabeled data are labeled at each repetition. While in Active-DECORATE, labeling is done manually by the user, our approach labels examples automatically by assigning them the labels (characteristics ratings, in the context of this research) with the highest probabilities/confidence as predicted by the current ensemble of classifiers.

Since the process of generating the ensemble of classifiers for every semantic characteristic is the same, we will explain below the general steps of our approach regardless of the semantic characteristic to be predicted. The only difference will consist of the initial labeled data that will be used for creation of the ensemble of classifiers. For each characteristic, the

ensemble will be built starting with the nodules on which at least three radiologists' agree with respect to that semantic characteristic (regardless of the other characteristics).

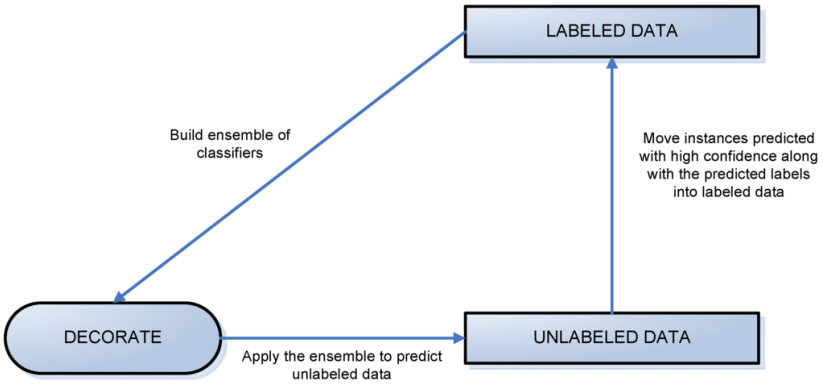

Figure 3. A diagram of the labeling process.

We divided the LIDC data into two datasets: labeled and unlabeled data, where labeled data included all instances of the nodules on which at least three radiologists agreed and unlabeled data contained all other instances (**Figure 3**). The algorithms woks iteratively to move all examples from the unlabeled data set to the labeled data set. At each iteration, some instances were chosen for this transition using the results of classification specific to that iteration.

Instances were added to the labeled data set based on the confidence with which they were predicted. Instances predicted with probability higher than a threshold were added into the training set along with their predicted labels (ratings produced by CAD). When an iteration of the algorithm failed to produce any labels of sufficient confidence, every instance left in the unlabeled pool was added to the labeled data along with its original label (rating assigned by the radiologist). This is shown by the vertical arrow in **Figure 3**. At this point, the ensemble of classifiers generated in the most recent iteration is the ensemble used to generate final classification and accuracy results.

The creation of the ensemble of classifiers at each iteration is driven by the DECORATE algorithm. The steps of the DECORATE algorithm are as follows: first, the ensemble is initialized by learning a classifier on the given labeled data. On subsequent steps, an additional classifier is learned by generating artificial training data and adding it to the existing training data. Artificial data is generated by randomly picking data points from a Gaussian approximation of the current labeled data set and labeling these data points in such a way that labels chosen differ maximally from the current ensemble's predictions. After a new classifier is learned based on the addition of artificial data, the artificial data is removed from the labeled data set and the ensemble checked against the remaining (original, non-artificial) data. The decision on whether a newly created classifier should be kept in the ensemble depends on how this classifier affects the ensemble error. If the error increases, the classifier is discarded. The process is repeated until the ensemble reaches the desired size (number of classifiers) or a maximum number of iterations are performed. A visual representation of the algorithm's steps is shown on present a visual overview of the ensemble of classifiers' agreement with the panel of experts' opinions. In this visualization, we were interested not only in the "absolute" accuracy of the classifier, but also in how the classifier did with regard to rater disagreement. For each semantic characteristic, we have displayed four graphs. Each one of these graphs corresponds to a distinct number of raters. That is, we show one graph for nodules rated by one radiologist (upper left graph in each figure), one graph for nodules rated by two radiologists (upper right graph in each figure), one graph for nodules rated by three radiologists (lower left graph in each figure) and one graph for nodules rated by four radiologists (lower right graph in each figure). In each graph, we have a bar corresponding to the number of radiologists which our algorithm predicted correctly. (Thus the graphs with more radiologists have more bars.) The height of the bars shows how many nodules there were in each level of prediction success. Looking at just the height of these bars, we can see that our classifier's success was quite good with respect to most of the semantic characteristics – these characteristics present very right-skewed distributions. Lobulation, spiculation and texture present more uniform distribution, meaning our classifier was less successful at predicting the radiologists' labels. We present one further visualization in these graphs–each bar is gray-coded to indicate the radiologists' level of agreement among themselves. (Thus, for example, the upper left graph, one radiologist, has no gray-coding, as a radiologist will always agree with himself.) This gray-coding allows us to see that the approach is much better at matching radiologists when the radiologists agree with themselves. While this, in itself, is not surprising, it does reveal that for the troublesome characteristics

(lobulation, spiculation and texture) the algorithm does a very good job when we look only at higher levels of radiological agreement.

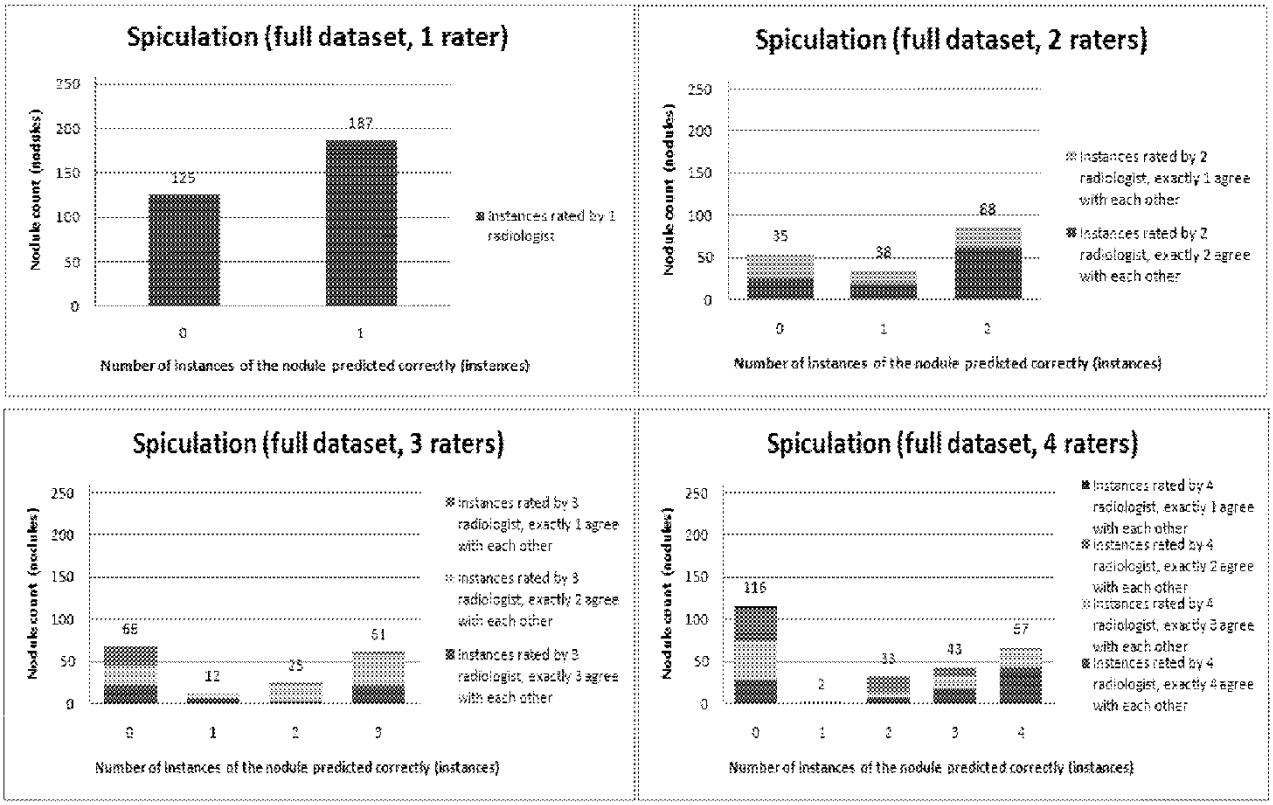

Figure 4. Visual overview of the ensemble of classifiers' agreement with the panel of experts' opinions (Spiculation).

5. Conclusions

In this paper, we presented a semi-supervised learning approach for predicting radiologists' interpretations of lung nodule characteristics in CT scans based on low-level image features. Our results show that using nodules with a high level of agreement as initially labeled data and automatically labeling the data on which disagreement exists, the proposed approach can correctly predict 70% of the instances contained in the dataset. The performance represents a 24% overall improvement in accuracy in comparison with the result produced by the classification of the dataset by classic decision trees. Furthermore, we have shown that using balanced datasets, our approach increases its prediction accuracy by 45% over the classic decision trees. When measuring the agreement between our computer-aided diagnostic characterization approach and the panel of experts, we learned that there is a moderate or better agreement between the two when there is a higher consensus among the radiologists on the panel and at least a 'fair' agreement when the opinions among radiologists vary within the panel. We have also found that high disagreement in the boundary delineation of the nodules also has a significant effect on the performance of the ensemble of classifiers.

In terms of future work, we plan to explore further (1) different classifiers and their performance with respect to the variability index in the expectation of improving our performance, (2) 3D features instead of 2D features so that we can include all the pixels in a nodule without drastically increasing the image feature vector size, and (3) integration of the imaging acquisition parameters in the ensemble of classifiers so that our algorithm will be stable in the face of images obtained from different models of imaging equipment. In the long run, it is our aim to use the proposed approach to measure the level of inter-radiologist variability reduction by supplying our CAD characterization approach in between the first and second pass of radiological interpretation.

References

1. Armato, S.G.; McLennan, G.; McNitt-Gray, M.F.; Meyer, C.R.; Yankelevitz, D.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; Kazerooni, E.A.; MacMahon, H.; Reeves, A.P.; Croft, B.Y.; Clarke, L.P. Lung Image Database

Consortium Research Group. Lung image database consortium: Developing a resource for the medical imaging research community. *Radiology* **2004**, *232*, 739–748

2. Raicu, D.; Zinovev, D.; Furst, J.; Varutbangkul, E. Semi-supervised learning approaches for predicting lung nodules semantic characteristics. *Intell. Decis. Technol.* **2009**, *3*, No. 2
3. Chapelle, O.; Schölkopf, B.; Zien, A. *Semi-Supervised Learning*; MIT: Cambridge, MA, USA, 2006
4. Cohen, J. Weighted kappa; nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220
5. McNitt-Gray, M.F.; Hart, E.M.; Wyckoff, N.; Sayre, J.W.; Goldin, J.G.; Aberle, D.R. A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results. *Med. Phys.* **1999**, *26*, 880–888
6. McNitt-Gray, M.F.; Wyckoff, N.; Sayre, J.W.; Goldin, J.G.; Aberle, D.R. The effects of co-occurrence matrix based texture parameters on the classification of solitary pulmonary nodules imaged on computed tomography. *Comput. Med. Imaging Graph.* **1999**, *23*, 339–348.
7. Lo, S.C.B.; Hsu, L.Y.; Freedman, M.T.; Lure, Y.M.F.; Zhao, H. Classification of lung nodules in diagnostic CT: An approach based on 3-D vascular features,nodule density distributions, and shape features. In Proceedings of SPIE Medical Imaging Conference, San Diego, CA, USA,, February, 2003; pp. 183–189.
8. Armato, S.G., III; Altman, M.B.; Wilkie, J.; Sone, S.; Li, F.; Doi, K.; Roy, A.S. Automated lung nodule classification following automated nodule detection on CT: A serial approach. *Med. Phys.* **2003**, *30*, 1188–1197.
9. Takashima, S.; Sone, S.; Li, F.; Maruyama, Y.; Hasegawa, M.; Kadoya, M. Indeterminate solitary pulmonary nodules revealed at population-based CT screening of the lung: using first follow-up diagnostic CT to differentiate benign and malignant lesions. *Am. J. Roentgenol.* **2003**, *180*, 1255–1263.
10. Takashima, S.; Sone, S.; Li, F.; Maruyama, Y.; Hasegawa, M.; Matsushita, T.; Takayama, F.; Kadoya, M. Small solitary pulmonary nodules (<1 cm) detected at population-based CT screening for lung cancer: reliable high-resolution CT features of benign lesions. *Am. J. Roentgenol.* **2003**, *180*, 955–964
11. Shah, S.; McNitt-Gray, M.; Rogers, S.; Goldin, J.; Aberle, D.; Suh, R.; DeZoysa, K.; Brown, M. Computer-aided lung nodule diagnosis using a simple classifier. *Int. Congr. Ser.* **2004**, *6*, 952–955
12. Holte, R.C. Very simple classification rules perform well on most commonly used datasets. *Mach. Learning* **1993**, *11*, 63–91
13. Samuel, C.C.; Saravanan, V.; Vimala, D.M.R. Lung nodule diagnosis from CT images using fuzzy logic. In Proceedings of International Conference on Computational Intelligence and Multimedia Applications, Sivakasi, Tamilnadu, India, December 13−15, 2007; pp. 159–163.
14. Sluimer, I.; Schilham, A.; Prokop, M.; Ginneken, B. Computer analysis of computed tomography scans of the Lung: A survey. *IEEE Trans. Med. Imaging* **2006**, *4*, 385–405
15. Goldin, J.G.; Brown, M.S.; Petkovska, I. Computer-aided diagnosis in lung nodule assessment. *J. Thoracic Imaging* **2008**, *23*, 97–104
16. Gurney, J. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part I. Theory. *Radiology* **1993**, *186*, 405–413.
17. Gurney, J.; Lyddon, D.; McKay, J. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part II. Application. *Radiology* **1993**, *186*, 415–422. [
18. Matsuki, Y.; Nakamura, K.; Watanabe, H.; Aoki, T.; Nakata, H.; Katsuragawa, S.; Doi, K. Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: Evaluation with receiver operating characteristic analysis. *Am. J. Roentgenol.* **2002**, *178*, 657–663.

19. Aoyama, M.; Li, Q.; Katsuragawa, S.; Li, F.; Sone, S.; Doi, K. Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. *Med. Phys.* **2003**, *30*, 387–394.

20. Kahn, C.; Channin, D.; Rubin, D. An ontology for PACS integration. *J. Digital Imaging* **2006**, *12*, 316–327.]

21. Barb, A.S.; Shyu, C.R.; Sethi, Y.P. Knowledge representation and sharing using visual semantic modeling for diagnostic medical image databases. *IEEE Trans. Inf. Technol. Biomed.* **2005**, *9*, 538–553.

22. Ebadollahi, S.; Coden, A.; Tanenblatt, M.A.; Chang, S.F.; Syeda-Mahmood, T.F.; Amir, A. Concept-based electronic health records: Opportunities and challenges. *ACM Multimed.* **2006**, 997–1006.

23. Ebadollahi, S.; Johnson, D.E.; Diao, M. Retrieving clinical cases through a concept space representation of text and images. *SPIE Med. Imaging Symp.* **2008**. (submitted

24. Nie, K.; Chen, J.H.; Yu, H.J.; Chu, Y.; Nalcioglu, O.; Su, M.Y. Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI. *Acad. Radiol.* **2008**, *15*, 1513–1525

25. Liney, G.P.; Sreenivas, M.; Gibbs, P.; Garcia-Alvarez, R.; Turnbull, L.W. Breast lesion analysis of shape technique: Semi-automated *vs.* Manual

26. Raicu, D.S.; Varutbangkul, E.; Cisneros, J.G.; Furst, J.D.; Channin, D.S.; Armato, S.G., III. Semantics and image content integration for pulmonary nodule interpretation in thoracic computed tomography. In Proceedings of SPIE Medical Imaging Conference, San Diego, CA, USA, February, 2007.

27. Raicu, D.S.; Varutbangkul, E.; Furst, J.D.; Armato, S.G., III. Modeling semantics from image data: opportunities from LIDC. *Int. J. Biomed. Eng. Technol.* **2008**, 1–22

28. Opulencia, P.; Channin, D.S.; Raicu, D.S.; Furst, J.D. Mapping LIDC, RadLex, and Lung nodule image features. *J. Digital Imaging* **2009**, (in press).

29. Cohen, J. A coefficient of agreement for nominal scale. *Educat. Psychol. Measure.* **1960**, *20*, 37–46.

30. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378–382.

31. Landis, J.R.; Koch, G.G. An application of hierarchical Kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **1977**, *33*, 363–374.

32. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174

33. Kraemer, H.C. Extension of the kappa coefficient. *Biometrics* **1980**, *36*, 207–216

34. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378–382.

35. Viera, A.J.; Garrett, J.M. Understanding interobserver agreement: The Kappa statistic. *Fam Med.* **2005**, *5*, 360–363. [

36. McNitt-Gray, M.F.; Armato, S.G., III; Meyer, C.R.; Reeves, A.P.; McLennan, G.; Pais, R.C.; Freymann, J.; Brown, M.S.; Engelmann, R.M.; Bland, P.H.; Laderach, G.E.; Piker, C.; Guo, J.; Towfic, Z.; Qing, D.P.; Yankelevitz, D.F.; Aberle, D.R.; van Beek, E.J.; MacMahon, H.; Kazerooni, E.A.; Croft, B.Y.; Clarke, L.P. The Lung image database consortium (LIDC) data collection process for nodule detection and annotation. *Acad. Radiol.* **2007**, *12*, 1464–1474.

37. Philips, C.; Li, D.; Furst, J.; Raicu, D. An analysis of Co-occurrence and gabor texture classification in 2D and 3D. In Proceedings of CARS, Barcelona, Spain; 2008.

38. Melville, P.; Mooney, R. Constructing diverse classifier ensembles using artificial training examples. In Proceedings of 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico; 2003; pp. 505–510.

39. Melville, P.; Mooney, R. Diverse ensembles for active learning. In Proceedings of International Conference on Machine Learning, Banff, Alberta, Canada, July, 2004; pp. 584–591.

40. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.

41. Siena, S.; Zinoveva, O.; Raicu, D.; Furst, J. Area and shape-dependent variability metric for evaluating panel segmentations of lung nodules in LIDC data. In Proceedings of SPIE Medical Imaging Conference, San Francisco, CA, USA, February, 2010. (accepted).