2011 10th International Conference on Machine Learning and Applications

Building an Ensemble of Probabilistic Classifiers for Lung Nodule Interpretation

Dmitriy Zinovev
College of Computing and Digital Media
DePaul University
Chicago, IL, US
dzinovev@cdm.depaul.edu

Jacob Furst
College of Computing and Digital Media
DePaul University
Chicago, IL, US
jfurst@cdm.depaul.edu

Daniela Raicu
College of Computing and Digital Media
DePaul University
Chicago, IL, US
dstan@cdm.depaul.edu

Abstract— **When examining Computed Tomography (CT) scans of lungs for potential abnormalities, radiologists make use of lung nodule's semantic characteristics during the analysis. Computer-Aided Diagnostic Characterization (CADc) systems can act as an aid - predicting ratings of these semantic characteristics to aid radiologists in evaluating the nodule and potentially improve the quality and consistency of diagnosis. In our work, we propose a system for predicting the distribution of radiologists' opinions using a probabilistic multi-class classification approach based on combination of belief decision trees and ADABoost ensemble learning approach. To train and test our system we use the National Cancer Institute (NCI) Lung Image Database Consortium (LIDC) dataset, which includes semantic annotations by up to four radiologists for each one of the 914 nodules. Furthermore, we evaluate our probabilistic multi-class classifications using a novel distance-threshold curve technique intended for assessing the performance of uncertain classification systems. We conclude that for the majority of semantic characteristics there exists a set of parameters that significantly improves the performance of the ensemble over the single classifier.**

Keywords-component; CAD; ensemble learning; uncertain classification; multi-class; belief decision trees

I. INTRODUCTION

Lung cancer is the most prevalent cause of cancer-related deaths in the US [1]. In order to effectively treat lung cancer early diagnosis of the disease has to be performed. In order to effectively diagnose a lung cancer, a radiologist analyses a series of computed tomography (CT) scans. The radiologist evaluates the development of the potential lung nodule as well as its visual properties to provide recommendations for the physician, which will be helpful in a diagnosis process. In order to improve the quality of such analysis a number of computer-aided diagnosis (CAD) systems have been recently developed. Such systems act as an aid in the evaluation process allowing increasing the quality of radiologist's recommendations and avoiding potential false positives and false negatives.

In order for the creator of such a system to evaluate its efficiency, some sort of reference standard dataset is necessary. An example of such a dataset is the Lung Image Database Consortium (LIDC) [2], a collection of CT studies analyzed by a panel of 4 radiologists. Each expert provided an outline for every nodule that he found in the dataset as well as the set of semantic ratings for that nodule. These characteristics are lobulation, malignancy, margin, sphericity, spiculation, subtlety, and texture and they were rated on a 5-point scale.

One of the properties of the LIDC dataset is the lack of ground truth data from biopsy or follow-up for the vast majority of LIDC nodules. As the radiologists were evaluating the nodules present in the dataset they were not forced to agree with each other and therefore, variability is present in both outlines and semantic ratings of different radiologists. Due to the 1) lack of information about the level of expertise of the different radiologists, 2) their anonymity across different nodules and 3) the lack of ground truth data there is no simple way to properly address this variability; therefore, the nodule is associated with a set of semantic ratings as opposed to a single rating. These challenges however, give the opportunity to apply non-traditional machine learning techniques to computer-aided diagnosis.

One of the most straight forward solutions for addressing variability in the interpretation is to artificially remove the variability by, for example, taking the mode opinion as a consensus rating of a lung nodule [3]. This approach has several drawbacks, including incorrect mode diagnosis in bimodal distributions of opinions or loss of potentially important information, when non-mode ratings are ignored [4].

In this paper we employ a different strategy for handling the variability by building classifiers able to learn from multi-

978-0-7695-4607-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICMLA.2011.44

class probabilistic labels and then combining these classifiers into an ensemble of classifiers. While in our previous work [5] we demonstrated that an ensemble of decision tree classifiers outperforms a single decision tree classifier trained on consensus-based label data, in this paper we investigate the same hypothesis but for distributions of diagnosis interpretations that will be used to create and validate belief decision trees. Furthermore, several adaptations are made to the belief decision trees and AdaBOOST to take into account the unbalanced nature of the LIDC data as most of the characteristics are strongly biased towards one of the ratings.

The rest of the paper is organized as follows: Section II discusses the related work in the area of multi-class and uncertain classification; Section III describes the dataset as well as the proposed methodology; Section IV presents the evaluation results, Section V discusses these results and Section VI summarizes our presented work and describes possible avenues for future work.

II. Related Work

The belief decision tree is a classification approach intended for learning from data with uncertain labels. The uncertainty can be due to the presence of multiple observers or uncertainty of the observer itself. The theoretical foundation of the algorithm was described by Elouedi et al. [6], in which the authors described the details of building a classifier using unlabeled data from a synthetic dataset with categorical objective features. Further, the algorithm was used for solving classification problems of different nature: Vannoorenberghe and Denœux [7] combined the algorithm with a one-versus-all technique to train a classification model, capable of classifying acoustic emission samples from data labeled by multiple observers. Trabelsi et al. [8] evaluated various methods of pruning the belief decision trees on various publically available datasets, two of them being from the medical domain. Elouedi et al. [9] employed the belief decision tree technique to assess the reliability of several jointly working sensors. Jenhani et al. [10] described a possibilistic belief decision tree classification algorithm, which was evaluated on several publically available datasets for which possibilistic labels were artificially generated.

Ensemble-based machine learning techniques are aimed at improving the performance of classification algorithm. Ensemble members are iteratively trained by introducing diversity into training data at every iteration of the ensemble learning process. One of the most popular ensemble learning techniques is ADABoost which is known to be a slowly overfitting algorithm and one of the best out-of-the-box ensemble learning approaches [11]. The algorithm was presented by Freund and Schapire in [12], in which the authors provided the theoretical justification, and discussed the potential applications of the proposed algorithm. The ADABoost technique has been widely used in the medical domain. Madabhushi et al. [13] employed ADABoost as a baseline for evaluating the performance of a CAD system that they developed for detecting prostatic adenocarcinoma. Ochs et al. [14] used ADABoost for classifying structures of lungs (nodules, airways, etc.) in CT images. Harirchi et al. [15] built a CAD system for automatic detection of micro calcifications in mammograms based on ADABoost. Quost and Denœux in [16] presented a creedal boosting algorithm based on ADABoost that was applied to the classification of two-class probabilistic data, including EEG signals.

Atif Tahie et al. [17] describe the heterogeneous ensemble learning technique RaKEL (Random k-Label sets) capable of building an ensemble of classifiers that were learned using different learning algorithms. The technique was intended for solving multi-label classification task. Authors evaluated the technique with various multi-label base classifiers using different multi-label evaluation metrics. Authors noticed the consistent performance boost for the ensemble vs. single classifier with respect to different evaluation techniques and training datasets.

In the Computer-Aided Diagnosis domain, the difference in the diagnostic interpretation has been addressed by either assessing the performance of each observer individually [18, 19] or by employing an “artificial consensus” upon the set of opinions [20, 21]. In the context of the LIDC dataset, the majority of CAD work reports systems for classifying the malignancy semantic characteristic based on agreement only [22, 23]. In our work we will address uncertainty, caused by the presence of multiple observers, by considering the whole range of opinions during the system training process and making use of the uncertain output labels produced by our classification system. Besides this uncertain approach, novel to the CAD medical domain, we will also investigate the whole range of the semantic properties (margin, texture, spiculation, sphericity, subtlety, and lobulation) of the lung nodules that were found to be important for the lung nodule diagnosis process. While we have recently looked into predicting the distributions of opinions in the radiological domain [4], to our best knowledge, there is no other work that combines ensembles of classifiers that emulate panels of experts with belief decision trees that predict the differences among their opinions.

III. Methodology

Our proposed methodology to handle the variability in the diagnosis process consists of several stages: first, image features are extracted from the nodules (Section A) to be interpreted and further used in the classification process. Second, belief decision trees classifiers (Section B) are build to predict the uncertainly labeled data with respect to each one of the seven semantic characteristics. Third, an ensemble of classifiers is constructed using an adaptation of the ADABoost approach (Section C) to test the hypothesis that an ensemble of classifiers significantly outperforms a single classifier. This third stage is the analog of having multiple experts involved in the interpretation process rather than a single one. As a final step, Area under Distance Threshold curve technique (Section D) is employed to evaluate the performance of probabilistic multi-class classifiers.

The proposed methodology is applied independently to each one of the seven semantic characteristics. Several considerations were taken into account when predicting each one of the characteristics individually. In the clinical environment, radiologists do not usually rate the malignancy of the nodule when providing the recommendations for the physician. Instead, they usually describe findings that they were able to identify as a suspicious mass with respect to the properties of this mass. Furthermore, previous work conducted in our lab has shown that the correlations between different semantic characteristics across different nodules were, in fact, very low. Lastly, when creating the LIDC dataset, radiologists were annotating each semantic characteristic without consideration of the ratings assigned to the other semantic characteristics, and therefore, the tasks of annotating different semantic characteristics were independent from each other. Taking these factors into account, we applied the methodology of this study individually to each one of the seven semantic characteristics.

A. LIDC dataset and nodule image features

The LIDC dataset (publically available from http://ncia.nci.nih.gov/) used in this research contains the CT images, radiologists' outlines of the lung nodules and subjective radiologists' semantic ratings on a scale from 1 to 5 for lobulation, malignancy, margin, sphericity, spiculation, subtlety, and texture.

The LIDC database currently contains complete thoracic CT studies for 399 patients acquired over different periods of time and with various scanners. 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 (Fig. 1.). Ground truth for the semantic ratings of lung nodules is not available for the LIDC dataset; therefore, ratings supplied by radiologists have to be used for training the classification system and evaluating the results.

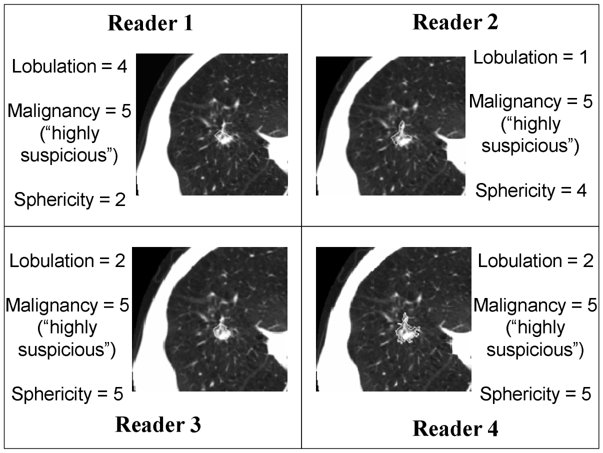
Reader 1
Lobulation = 4
Malignancy = 5 ("highly suspicious")
Sphericity = 2
Reader 2
Lobulation = 1
Malignancy = 5 ("highly suspicious")
Sphericity = 4
Lobulation = 2
Malignancy = 5 ("highly suspicious")
Sphericity = 5
Reader 3
Lobulation = 2
Malignancy = 5 ("highly suspicious")
Sphericity = 5
Reader 4

Figure 1. Visual representation of the LIDC data structure; one nodule is exemplified through the differences in the nodule's outlines and semantic ratings.

In this study we considered 914 nodules greater than 5×5 pixels in size for which we calculated a set of 63 two-dimensional image features from four categories: shape features, texture features, intensity features, and size features. The details of feature extraction process are described in our previous work [5]. For every nodule we considered only a single slice where the area of the nodule was largest with respect to up to 4 outlines provided by the radiologists who annotated the nodule; therefore for each nodule, a set of image features was calculated from single slice only. After extraction of the features, feature vector was concatenated with 5-class probability distribution constructed from semantic annotations by 4 radiologists to create a vector representation of a nodule.

B. Belief decision trees

As probabilistic base classifier for this research we implemented and adapted the classification approach proposed by Elouedi et al. [6]. Classification is performed in a manner similar to the one of regular decision trees. At 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 one with respect to the information gain that the split produces 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 a certain number of instances that reached this node and only twice that many that reached its parent (5 and 10 were determined empirically as a compromise between complexity of the classification model and training dataset cross-validation performance). The change in the number of instances from the original algorithm prevents our approach to do an overfitting.; 2) all BBAs of the instances which reached the node are equal; 3) all the available attributes/features are used for splitting; or 4) the information gain of all possible further splits is equal to 0.

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

First, the algorithm 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:

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

Where C is a belief mass that C_i is a member of Θ, Θ is a set of all possible classes and $m^{\Theta}\{I_j\}(C)$ is a probability associated with the corresponding belief mass C and $m^{\Theta}\{I_j\}(0)$ is a probability associated with the belief mass of an instance not being a member of any class from the available pool of classes. Due to the fact that all BBAs in the LIDC dataset are singletons (each radiologist had to pick one class and one class only when assigning the rating to a nodule), 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).

$$BetP^{\Theta}\{I_j\}\{C_i\} = \frac{\lambda_i}{\sum_{l=1}^{5} \lambda_l} \tag{2}$$

(where λl={0,1,2,3,4} is the rater count for every class C_i rated on a scale from 1 to 5)

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

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

Third, it computes the entropy of average pignistic probabilities in S:

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

where n is the number of possible classes.

For every attribute/feature, the algorithm creates a set of split threshold values in such a way that every distinct pair of values in the sorted set of attribute values produces a separate threshold. Next, for each of the thresholds, the algorithm collects two subsets S_V^A made with the cases having V as a value below the certain threshold – for the first subset and above the certain threshold – for the second subset for the attribute A, and computes the pignistic probability $BetP^{\Theta}\{S_V^A\}\{C_i\}$ for every class for each of two subsets for attribute A (equation 3). Finally the algorithm computes $Info_A(S)$ for every attribute as:

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

Where $Info(S_V^A)$ is calculated using equation (4).

The original algorithm proposed by Elouedi et al. [6] was described for the categorical instance's attributes, which allowed the model to produce a small number of natural splits. Since there is no best way to pick a split value for the numerical attribute, a step that tests multiple splitting threshold for goodness of split was introduced.

To calculate goodness of split, the algorithm computes the information gain:

$$Gain(S, A) = Info(S) - Info_A(S) \tag{6}$$

The combination of attribute/feature and split threshold value for this attribute that produced the largest value of the information gain is used for the split. Information gain criteria was used in our approach to determine goodness of split due to the fact that gain ration criteria used by algorithm described in [6] produced very unbalanced splits at every step of model training. Such unbalanced splits created terminal nodes with very small subsets of instances, which could potentially lead to overfitting of the classification model.

C. *ADAboost ensemble learning algorithm*

In this paper we chose the ADABoost algorithm as an ensemble learning approach to create an ensemble of probabilistic classifiers using belief decision tree as a base classifier. ADABoost is known as an algorithm that aims at creating a combination of weak learners that together will act as a strong learner and improve the classification performance over a single classifier trained on the same set of instances. The training of the ensemble model is carried out as follows:

Given a training set S of instances = [(*I_j*, *y_j*)], *j* = 1,…,|*S*| with probabilistic labels $y_j \in BetP^{\Theta}\{I_j\}\{C_1\}$,..., $BetP^{\Theta}\{I_j\}\{C_n\}\}$ ADABoost creates an ensemble of classifiers H as follows:

In the initial step it assigns equal weights D_1(j) to every training instance I_j and normalizes the weights to convert them into probabilities: $D_1(j) = 1/|S|$

For t iterations from 1 to T (where T is the desired number of ensemble members) the algorithm will:

- Replicate instances in a training dataset proportionally to the normalized weights of instances, thus ensuring that for every member of the ensemble of classifiers, each training instance participates in the training process at least once (Fig. 2).
- Train a classifier h_t on the created subset S_t.
- Classify every instance of the training dataset and calculate h_t classifier's error ε_t as a sum of weights of misclassified instances. Classifier h_t is kept in the ensemble if $0 < \varepsilon_t < 0.5$

$$h_t \colon \varepsilon_t = \sum_{j:\arg(\mathrm{Max}(h_t(I_j)))\neq\arg(\mathrm{Max}((y_j))} D_t(j) \tag{7}$$

Where the arg function represents the argument for which the maximum probability is reached for the predicted BBA probabilities $h_t(I_j)$ and the radiologists ones, respectively.

- Calculate weight of the classifier as error/(1-error)

$$\beta_t = \frac{\varepsilon_t}{(1-\varepsilon_t)} \tag{8}$$

- Recalculate weights of the training instances by multiplying the weights of misclassified instances by classifier weight and renormalizing the weights distribution across the training dataset.

$$D_{t+1}(j) = \frac{D_t(j)}{Z_t} *$$

$$\begin{cases} 1\ if\ \arg(Max((h_t(I_j))) = \arg(Max((y_j)) \\ \qquad \beta_t\ otherwise \end{cases} \tag{9}$$

$$Z_t = \sum_j D_t(j) \tag{10}$$

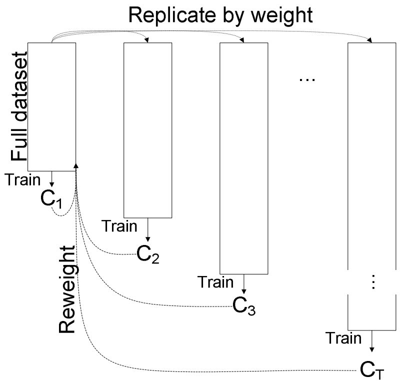

Figure 2. Process of replicating instances by weights; new ensemble member is trained on the whole training dataset, where instances misclassified at the previous iteration are replicated proportionally to their weights.

After the ensemble is created it is evaluated on entire training, testing and validation datasets by taking a weighted average of probabilities produced by the ensemble members, where weights are based on the weights of the ensemble members. $\mathrm{H}(\mathrm{I}_\mathrm{j})$ is the ensemble's discrete probability density function over the label set $C_1, \dots, C_n$:

$$\mathrm{H}(\mathrm{I}_\mathrm{j}) = \frac{\sum_{t=1}^{T} \mathrm{h}_\mathrm{t}(\mathrm{I}_\mathrm{j}) * \log\left(\frac{1}{\beta_t}\right)}{|T|} \tag{11}$$

D. Performance Evaluation

When evaluating a classification system that utilizes a probability distribution of 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 [25], to assess the performance of a multiple-label classification approach. We were not able to construct ROC curve for the results that we obtained since the definitions of true positive rate and false positive rate are not directly applicable to the multiple-label classification task.

The distance curve is defined as follows:

Let y be a sequence of instance labels, $y= [y_1,y_2,\dots y_j\dots y_{|S|}]$ where |S| is the number of instances and each y_j is a discrete probability density function over the label set.

Similarly, let H(I) be a sequence of predicted labels, P = $[H(I_1),H(I_2),\dots H(I_j)\dots H(I_{|S|})]$ where each $H(I_j)$ is discrete probability density function over the label set.

Let Dist be a normalized distance function defined on the instance/prediction pairs, Dist $(y_j,H(I_j)) \in [0,1]$. We define the distance-threshold curve as

$$\frac{\sum_{j=1}^{N} [\mathrm{Dist}\,(y_j, H(I_j)) \leq x]}{N} \tag{12}$$

where x, threshold value for the distance, is defined from 0 to 1, and the [] are Iverson brackets, which equal 1 when the statement inside the brackets is true and 0 otherwise. It can be seen that values of the curve itself are between 0 and 1 and that the curve is monotonically increasing.

We define the area under the distance-threshold curve simply as

$$\int_o^1 \frac{\sum_{j=1}^{|S|} [Dist(y_j, H(I_j)) \leq x]}{|S|}\, dx \tag{13}$$

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, with 0 being a threshold value, 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 ($\mathrm{AuC}_{\mathrm{dt}}$) was used as the metric for comparison. For this study, we used the Jeffrey Divergence distance metric [26] to generate the Dist distance function for formula (12).

IV. Results

In order to evaluate the performance of the proposed approach the available dataset of lung nodules was divided as follows. First, 10% of all instances were sampled from the dataset in such manner that the distribution of labels in the 10% subset would mimic the distribution of labels in the remaining 90%. This 10% dataset is used as validation data or reserved data, unseen by the classification algorithm until the moment the model was created. At the second step, the 90% dataset was divided into two parts, containing 66% and 34% of the remaining instances respectively. The division was done in a same manner as before, preserving the original label distribution in both resulting subsets. 66% subset was used for training the model and 34% subset was used for testing it. We created 3 different training/testing splits and further used them for training and testing across different experimental setups to make performance values directly comparable with each other. Even though the 66%/34% split was different at every time leading to the overlap between testing and training subsets which could potentially bias the calculated performance, 10% validation subset remained the same for every split, providing more reliable performance measure.

The ensemble of classifiers was created with the following settings: the ensemble consisted of 10 members; complexity of the individual trees defined by the number of instances at the parent of a terminal node was tested with 10, 20, 25 and 30 instances, the ensemble membership performance threshold used to decide whether to keep a classifier in the ensemble was set to 50%. Overall, for both single belief decision trees and the ensemble approach, the experiment was repeated 12 times, with 4 different levels of tree complexity (number of instances at the parent of the terminal node) on each of 3 fixed trials.

The performance results of the ensemble approach for the sphericity semantic characteristic are not reported due to the fact that during the ensemble building process the performance of newly created members converged below the ensemble membership accuracy threshold.

Table I reports classification performance of the single Belief Decision tree classifier on training, testing and validation subsets.

TABLE I SINGLE BDT

	Tree Complexity	Training 66		Testing 34		Reserved 10	
		ACC	AuCdt	ACC	AuCdt	ACC	AuCdt
Lobulation	30	66.9	73.3	57.4	74.7	63.4	74.1
Malignancy	10	60.0	73.0	48.8	65.0	55.9	70.4
Margin	10	58.3	72.3	53.8	64.3	47.6	64.6
Spiculation	25	70.2	81.4	70.3	78.4	77.1	81.2
Subtlety	20	55.0	63.4	40.2	65.8	44.5	66.6
Texture	25	75.9	78.7	71.0	75.1	75.1	80.9
AVG	**20**	**64.4**	**73.7**	**56.9**	**70.5**	**60.6**	**73.0**

Table II reports the classification performance of ADABoost with BDTs for the parameters described above. The performance values in both tables are averaged across 3 different training/testing splits and reported for the tree complexity level (number of instances at the parent node) that produced the best accuracy for each one of the semantic characteristics.

TABLE II ADABOOST

	Tree Complexity	Training 66		Testing 34		Reserved 10	
		ACC	AuCdt	ACC	AuCdt	ACC	AuCdt
Lobulation	20	69.4	80.6	62.3	77.7	73.4	80.7
Malignancy	25	63.1	74.6	52.9	64.0	61.3	67.9
Margin	10	71.4	77.7	47.9	70.0	39.7	62.0
Spiculation	30	73.3	83.2	71.6	78.8	74.2	85.4
Subtlety	25	60.1	79.5	46.4	63.8	54.8	71.9
Texture	25	81.7	82.3	75.9	78.5	80.0	80.5
AVG	**22.5**	**69.8**	**79.7**	**59.5**	**72.1**	**63.9**	**74.7**

Table III reports the p-values for the tests conducted to determine whether the classification performances of the single classifier and ADABoost ensemble on 10% reserved subset were statistically significant.

TABLE III P-VALUES FOR SIGNIFICANCE TESTS ON DIFFERENCES IN PERFORMANCE OF SINGLE BDT AND ADABOOST ON RESERVED 10 SUBSET OF INSTANCES

	p-value (ACC)	p-value(AuCdt)
Lobulation	0.018	0.035
Malignancy	0.058	0.09
Margin	0.035	0.089
Spiculation	0.081	0.056
Subtlety	0.020	0.054
Texture	0.054	0.119
AVG	**0.081**	**0.098**

V. DISCUSSION

When comparing the average performance across 6 semantic characteristics we noticed that the performance boost when using the ensemble is statistically significant with respect to both accuracy and area under the curve for training dataset (5.4% accuracy and 6% area under the curve improvements), testing dataset (2.6% accuracy and 1.6% area under the curve improvements) and validation dataset (3.3% accuracy and 1.7% area under the curve improvements).

When examining the accuracy performance per semantic characteristic we determined that for 4 characteristics out of 6, the ensemble of classifiers significantly outperforms a single belief decision tree classifier with respect to accuracy on reserved subset. The significance testing was carried out at a 0.1 level of significance. These were the configurations that were reported in Table I and Table II.

The semantic characteristics of lung nodule are independent of each other, therefore predicting each of those characteristics is a separate classification problem. Given that the distribution of ratings, dominant rating and relation to the low-level image features are different for different semantic characteristic, inconsistency in the parameters set for training the optimal classification model, across different semantic characteristics is acceptable for our task.

VI. CONCLUSION AND FUTURE WORK

In this paper we proposed an adaptation of the ensemble learning ADABoost classification approach based on replication instead of a 50% sampling needed for unbalanced datasets. Furthermore, we combined the adapted ADABoost approach with belief decision trees to address the uncertainty in the diagnosis process captured through a distribution of opinions rather than consensus. Our results show that it is possible to model radiologists' interpretation variability and that an ensemble of classifiers boosts the performance over single classifiers for the majority of the semantic characteristics. Our next steps will be to investigate the construction of the ensemble of classifiers based on an area under the curve threshold instead of accuracy given that the later forces the algorithm to look at the mode of the probabilistic labels when adding a classifier in the ensemble rather than considering their distributions as for the area under the curve threshold. Furthermore, we plan to employ ensemble of classifiers based on semi-supervised active learning techniques in order to take advantage in the training process of the cases on which there is agreement among radiologists.

REFERENCES

[1] Cancer Facts and Figures, American Cancer Society, 2010.

[2] S. G. Armato 3rd, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, A. P. Reeves, B. Y. Croft and L. P. Clarke , "Lung Image Database Consortium: developing a resource for the medical imaging research community," Radiology, vol. 232, 2004, pp. 739-748.

[3] R. Ochs, H.J. Kimb, E. Angel, C. Panknin, M. McNitt-Gray, and M. Brown, "Forming a reference standard from LIDC data: impact of reader agreement on reported CAD performance," in Proceedings of the SPIE, vol. 6514, pp. 65142A-1–65142A-6, March 2007.

[4] D. Zinovev, J. Feigenbaum, J. Furst, and D. Raicu, "Probabilistic Lung Nodule Classification with Belief Decision Trees", 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2011), in press.

[5] D. Zinovev, D. Raicu, J. Furst, and S. Armato, "Predicting Radiological Panel Opinions using a Panel of Machine Learning Classifiers", Algorithms 2009, 2, pp. 1473-1502.

[6] Z. Elouedi, K. Mellouli, and P. Smets, “Belief decision trees: Theoretical foundations,” Int. J. Approx. Reason, vol. 28, 2001, pp. 91–124.

[7] P. Vannoorenberghe and T. Denœux, “Handling uncertain labels in multiclass problems using belief decision trees,” Proceedings of the 5th IPMU Conference, pp. 1919–1926, Annecy, France, July 2002.

[8] S. Trabelsi, Z. Elouedi, and K. Mellouli. “Pruning belief decision tree methods in averaging and conjunctive approaches,” Int. J. Approx. Reasoning 46, 3, pp. 568-595, December 2007.

[9] Z. Elouedi, K. Mellouli, and P. Smets, “Assessing Sensor Reliability for Multisensor Data Fusion within the Transferable Belief Model,” IEEE Transactions on Systems, Man & Cybernetics, Part B, 34(1), pp. 782–787, February 2004.

[10] I. Jenhani, N. Ben Amor, S. Benferhat, and Z. Elouedi, “SIM-PDT: a similarity based possibilistic decision tree approach,” in Proceedings of the 5th international conference on Foundations of information and knowledge systems, Berlin, Heidelberg, pp. 348-364, February 2008.

[11] P. Bühlmann and T. Hothorn, “Boosting algorithms: regularization, prediction and model fitting (with discussion)”. Statistical Science 22, pp. 477-522.

[12] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in Proceedings of the Second European Conference on Computational Learning Theory, Springer-Verlag, London, UK, pp. 23-37, 1995.

[13] A. Madabhushi, M. D. Feldman, D. N. Metaxas, J. Tomaszeweski, and D. Chute, “Automated detection of prostatic adenocarcinoma from high-resolution ex vivo MRI,” IEEE Transactions on Medical Imaging, 24(12), pp. 1611-1625, December 2005.

[14] R. A. Ochs, J. G. Goldin, F. Abtin, H. J. Kim, K. Brown, P. Batra, D. Roback, M. F. McNitt-Gray and M.S. Brown , “Automated classification of lung bronchovascular anatomy in CT using AdaBoost,” Medical Image Analysis, Volume 11, Issue 3, pp. 315-324, June 2007.

[15] F. Harirchi, P. Radparvar, H. A. Moghaddam, F. Dehghan, M. Giti, “Two-Level Algorithm for MCs Detection in Mammograms Using Diverse-Adaboost-SVM,” in Proceedings of ICPR'2010, Istanbul, Turkey, pp. 269-272, August 2010.

[16] B. Quost and T. Denœux, “Learning from data with uncertain labels by boosting credal classifiers,” in Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data, ACM, New York, USA, pp. 38-47, 2009.

[17] M. Atif Tahir, J. Kittler, K. Mikolajczyk, and F. Yan, "Improving Multilabel Classification Performance by Using Ensemble of Multi-label Classifiers", Multiple Classifier Systems (MCS), LNCS 5997, pp. 11–21, 2010.

[18] Y. Matsuki, K. Nakamura, H. Watanabe, T. Aoki, H. Nakata, S. Katsuragawa, and K. Doi, “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., vol. 178, no. 3, pp. 657–663, 2002.

[19] F. Li, M. Aoyama, J. Shiraishi, H. Abe, Q. Li, K. Suzuki, R. Engelmann, S. Sone, H. MacMahon and K. Doi, “Radiologists’ performance for differentiating benign from malignant lung nodules on high-resolution CT using computer estimated likelihood of malignancy,” AJR, 183, pp. 1209–1215, 2004.

[20] J. W. Fletcher, S. M. Kymes, M. Gould, N. Alazraki, R. E. Coleman, V. J. Lowe, C. Marn, G. Segall, L. A. Thet, K. Lee , “A comparison of the diagnostic accuracy of 18FFDG PET and CT in the characterization of solitary pulmonary nodules,” J Nucl Med, 49, pp. 179–85, 2008.

[21] S. G. Armato 3rd, R. Y. Roberts, M. Kocherginsky, D. R. Aberle, E. A. Kazerooni, H. Macmahon, E. J. van Beek, D. Yankelevitz, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, P. Caligiuri, L. E. Quint, B. Sundaram, B. Y. Croft, L. P. Clarke , “Assessment of radiologist performance in the detection of lung nodules: Dependence on the definition of “truth”,” Academic Radiology 16, pp. 28–38, 2009.

[22] S. C. B. Lo, L. Y. Hsu, M. T. Freedman, Y. M. F. Lure, H. Zhao, “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, pp. 183–189, February, 2003.

[23] S. Takashima, S. Sone, F. Li, Y. Maruyama, M. Hasegawa, M. Kadoya, “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, pp. 1255–1263.

[24] J. R. Quinlan, “Improved Use of Continuous Attributes in C4.5,” J Artif Intell Res, vol. 4, 1996, pp 77-90.

[25] K. A. Spackman, “Signal detection theory: Valuable tools for evaluating inductive learning,” Proc. 6th Int. Workshop on Machine Learning, pp. 160–163, 1989.

[26] H. Liu, D. Song, S. Rüger, R. Hu, V. Uren, “Comparing dissimilarity measures for content-based image retrieval,” Proc. 4th Asia Inf. Ret. Conf. on Information Retrieval Technology, pp. 44-50, January 2008.