Building confidence and credibility into CAD with belief decision trees

Conference Paper · March 2017

DOI: 10.1117/12.2255559

5 authors, including:

Rachael Affenit
Illinois Institute of Technology
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Jacob D. Furst
DePaul University
53 PUBLICATIONS **509** CITATIONS
SEE PROFILE

Daniela Stan Raicu
DePaul University
123 PUBLICATIONS **1,398** CITATIONS
SEE PROFILE

Building confidence and credibility into CAD with belief decision trees

Rachael N. Affenit[2], Erik R. Barns[1], Jacob D. Furst[1], Alexander Rasin[1], Daniela S. Raicu[1]
[1] College of Computing and Digital Media, DePaul University, Chicago, IL, US
[2] College of Science, Illinois Institute of Technology, Chicago, IL, US

ABSTRACT

Creating classifiers for computer-aided diagnosis in the absence of ground truth is a challenging problem. Using experts' opinions as reference truth is difficult because the variability in the experts' interpretations introduces uncertainty in the labeled diagnostic data. This uncertainty translates into noise, which can significantly affect the performance of any classifier on test data. To address this problem, we propose a new label set weighting approach to combine the experts' interpretations and their variability, as well as a selective iterative classification (SIC) approach that is based on conformal prediction. Using the NIH/NCI Lung Image Database Consortium (LIDC) dataset in which four radiologists interpreted the lung nodule characteristics, including the degree of malignancy, we illustrate the benefits of the proposed approach. Our results show that the proposed 2-label-weighted approach significantly outperforms the accuracy of the original 5-label and 2-label-unweighted classification approaches by 39.9% and 7.6%, respectively. We also found that the weighted 2-label models produce higher skewness values by 1.05 and 0.61 for non-SIC and SIC respectively on root mean square error (RMSE) distributions. When each approach was combined with selective iterative classification, this further improved the accuracy of classification for the 2-weighted-label by 7.5% over the original, and improved the skewness of the 5-label and 2-unweighted-label by 0.22 and 0.44 respectively.

Keywords: Computer-aided diagnosis, LIDC, belief decision tree, iterative classification, conformal prediction, confidence

1. INTRODUCTION

Early detection and diagnosis is critical in cases of potential lung cancer. Misdiagnosis of a malignant lesion can significantly impact a patient's chances of survival. Small malignant nodules are more frequently missed on CT radiographs, simply due to the difficulty of visually detecting them in the image.[1] Computer-aided diagnosis (CAD) systems can be used to analyze medical images and provide a "second opinion" for diagnosis. A CAD system consists of two components: a feature extraction component in which automatic low-level image features (such as shape, size, and texture) are automatically extracted from pixel image data and a classification component that classifies a lesion with respect of its degree of malignancy using the low-level image features. CAD systems can help the radiologist detect and diagnose small lesions (such as lung nodules) using objective image features quantifying lesion content.

The classifiers involved in these systems can be powerful predictors, but only if there is an effective sample ground truth to train them on, one which should cover the range of cases we can generally expect to encounter in a clinical setting. Availability of ground truth for training the CAD systems is challenging because it may require pathology reports which are not available for all cases as is the case when the medical images are obtained during screening. In the absence of ground truth, a reference truth based on an expert of panel of experts' opinions can be formed and used to train these CAD classifiers. Several studies showed substantial observer variability not only among non-specialized radiologists during standard clinical reporting[2], but also at the expert level of image interpretation[3]. Although there are several CAD studies[4-6] that looked at the performance of individual radiologists before and after using CAD, most CAD systems consider consensus as the reference standard for development and evaluation when interpretations of multiple radiologists are available. Within the consensus approach, either only the consensus opinion is known[7] or individual interpretations are known but a consensus opinion is formed[1,8]. However, the use of a consensus standard does not show the extent to which a new technique is accurate in establishing a diagnosis, but only the extent to which a new technique agrees with the consensus reading. This becomes problematic since although a consensus might be reached, nothing guarantees that the consensual result is true, as mere consensus does not imply the correctness of a diagnostic decision[9]. Therefore, we propose to build classification system for learning from and predicting the whole or a weighted version of the distribution of radiologists' annotations. As we pointed out in our previous work[10], such approach can be beneficial for CAD purposes for several reasons: learning from the distribution of annotations will help to avoid the loss of potentially important information when classification system has no knowledge on radiologists' level of expertise; probabilistic prediction can carry important information besides the malignancy of the nodule.

Recent work in the machine learning community tackle the problem of annotator variability by outputting measures of prediction reliability, which can be calculated by incorporating conformal prediction (CP) into the classifier. Conformal prediction is a method of determining the reliability of a classifier's prediction and its implementation involves the use of a new "calibration" set to determine how well a case prediction conforms to its actual classification, which is then compared with the conformity between the predicted test case distributions and each possible classification to determine the confidence and credibility of a prediction. In other words, the confidence and credibility values represent the likelihood that the predicted label is right and that the remaining labels are wrong[11], respectively. In this paper, we are investigating how the incorporation of these reliability measures into the classification process, can help with the robustness and reliability of the CAD systems in presence of uncertain diagnostic label data. More specifically, we propose to train a selective iterative classifier (SIC) that starts its training with one label and keeps adding another label for each case at each iteration only if the confidence of the prediction for the correspondent case is below a certain threshold; the process stops when the confidence of the classifier is above the threshold or there are no labels left to add.

Using the NIH/NCI Lung Nodule Image Database[12] (LIDC), we show an improvement in accuracy when predicting lung nodule malignancy by using conformal prediction and the selective iterative approach as well as when integrating a scheme for weighting labels based on their probabilistic distribution. Providing measures of confidence and reliability for the CAD output can both improve performance and increase radiologists' trust in CAD systems which can lead to a large-scale adoption of the CAD systems in the radiologists' workflow.

2. RELATED WORK

While many CAD systems have been built for different anatomic structures and types of lesions, only few provide additional built-in performance evaluation features to complement the CAD prediction outcome. We name the CAD systems that provide an explanation of their output in terms of confidence and/or reliability as well as contextual information to better inform diagnosis as Smart CAD systems. Examples of such systems include the ones that have been developed for breast and biliary structures.

Drukker, et. al.[13] developed a CAD system for breast ultrasound, which examined the difference between calculated nodule boundaries and radiologist marked nodule boundaries to determine confidence in diagnosis. They also examined different uses for the confidence measure, including as output to the radiologist, or as a self-evaluation measure to assess the performance of specific classifiers among multiple classifiers. Jagdale, et. al.[14] developed a CAD system for mammography which used a Bayesian network classifier to distinguish tumor cells from healthy tissue. Their system was designed to output an image of the cancerous tissue at the location of the region of interest (ROI), as well as the coordinates and size of the region. Marzieh et. al.[15] developed a system called Smart Atlas to identify biliary structures from confocal laser endomicroscopy (pCLE) video. They chose to include measures of specificity and positive predicted value (PPV) in their output to provide contextual information to the medical professional. Johannson, et. al.[11] described a mathematical method of integrating conformal prediction with decision trees; conformal prediction was also used by Harris et. al.[16] in conjunction with a CAD system for acute abdominal pain that used neural networks as the base classifier.

Handling the uncertainty of the medical label data has also been investigated either through iterative classification or probabilistic distribution encoding of the labels. Selective iterative classification builds R separate models, each trained on a different number of ratings from the training data, where R is the maximum number of ratings/annotations available. When new cases are being classified, they are first run through the model built with one rating. If this model produces a satisfactory prediction, this prediction is final, and no further models are used for classification. If this prediction is unsatisfactory, the case is classified by the next model, and the process repeats. If none of the models produces a satisfactory prediction, then some aggregate prediction of the r models is taken, usually an average across all the models. Riely et al.[17] showed that by using selective iterative classification and removing radiologist ratings that produced "noise" in the LIDC dataset, the running time of their classifier improved as well as its accuracy. Rather than attempting to produce a consensus label, Zinovev et. al.[18] proposed the use of belief decision trees[3] to produce probabilistic classifications for cases that had the label encoded with vectors of possibilities; they showed that using belief decision trees results in better accuracy for predicting malignancy than when using deterministic labels.

3. METHODOLOGY

We build on our previous work by investigating the merits of integrating selective iterative classification with belief decision trees and using conformal prediction to offer a self-correcting feature for the CAD system. As part of the methodology, we also investigate the impact of reducing and weighting the label set on the classification results by introducing a new label weighting approach that can also overcome the effects of majority voting we outlined in our previous work[19]. We demonstrate our approach on the NIH/NCI LIDC dataset in which the malignancy was rated on a scale from 1 to 5 (with 1 being "most likely benign" and 5 being "most likely malignant") and by up to four radiologists.

3.1 LIDC dataset

The publicly available LIDC database[12] (downloadable through the National Cancer Institute's Imaging Archive web site - http://ncia.nci.nih.gov/) provides Computed Tomography (CT) image data, nodules' outlines as delineated by radiologists, and the radiologists' subjective ratings of nodule characteristics (with respect to lobulation, malignancy, margin, sphericity, spiculation, subtlety, and texture). The LIDC database contains complete thoracic CT studies for about 1,000 patients acquired over different periods of time and with various scanners on which approximately 2,600 nodules were outlined and annotated. Each study can contain several nodules of a different size; therefore, there may be a different number of slices associated with a particular nodule. Ground truth for the semantic ratings of lung nodules is not available for LIDC dataset, therefore ratings supplied by radiologists have to be used for training the classification systems and evaluation of the results. The agreement among the four radiologists with respect to the malignancy diagnosis happens in only 25% of these cases[20] and thus, the analysis of the uncertain label data is well motivated by the LIDC dataset. We select a subset of the LIDC data which includes only those 809 cases where all four radiologists identified and rated a nodule. This subset is further balanced by under sampling the uncertain (rating of 3) cases by removing ~150 cases, and over sampling cases with each of the other labels by randomly duplicating ~50 cases, for a final balanced set of 850 cases.

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, we are considering only the slice in which the nodule has the largest area with respect to the outlines provided by up to four radiologists who annotated the corresponding nodule. Therefore, only the largest outline is considered as the most representative for feature extraction. After completion of the feature extraction process, we create a vector representation of every nodule which consists of 63 image features and 7 radiologist annotations. More details on the feature calculations and the rating values for each semantic characteristic are provided in Zinovev et al[21].

3.2 Belief decision trees

The construction of the belief decision trees is similar to the traditional decision trees with the exception that the decision to split a node is based on averaging the pignistic probabilities (called basic belief assignment (BBA)) for all the cases that reach that node of the tree. In the case of the LIDC data, the BBA uses the set of probabilities for the five malignancy labels to calculate the average BBA of all cases in the training set that reached a specific node[22]; for example, a rating distribution of [2, 3, 4, 4] would yield the BBA [0 .25 .25 .5 0] for a five-label distribution. The process of calculating these probabilities is typically more complex for a belief decision tree, but the LIDC dataset has some characteristics that allow us to use simplify this method. More specifically, since every radiologist can only choose one malignancy rating for each case, we can eliminate the possibility of having two or more "true" labels. The LIDC dataset also has no representation of pure uncertainty (as a rating of 3 indicates balanced probabilities of malignant or benign labels), resulting in a simplification of this calculation to a probability distribution[18].

The average pignistic probabilities of the parent and child nodes can then be used to calculate the information gain to decide the split with respect to each possible feature and threshold value in the dataset. Given that the information gain split criteria can produce heavily unbalanced trees, we propose to use the gain ratio which controls for the size of the child subsets and rewards equally distributed splits. The low-level image feature that achieves the maximum gain ratio for the corresponding split is then selected at that node to perform the partitioning of the cases. Furthermore, to determine whether a node in a BDT is a leaf node, one could use one of four stopping criteria: the maximum gain ratio of splitting is 0, there is no split that can be made which will result in acceptable numbers of cases at the parent and child nodes (given by n_p and n_c parameters, respectively), all the BBA's at the node are equivalent, or all features have already been used to split[22].

3.3 Conformal prediction

We propose to combine the BDT classifier with Conformal Prediction (CP) to produce measures of confidence and credibility for each CAD probability distribution output. Conformal Prediction begins as a typical classification problem: the dataset is divided into a training and a testing set, but the training set must be further divided into a proper training set and a calibration set. For our implementation, we define the calibration set as a randomly selected 1/7 of the training set, and the proper training set as the remaining cases from the training set. This ratio was experimentally determined on our balanced dataset, as any larger of a calibration set would not leave enough cases to satisfactorily train our classifier. The calibration set is used to facilitate conformal prediction by providing a base set of conformity scores.

The calibration set is classified using the BDT produced by the training set, and the conformity function given in equation 1 is used to determine conformity scores α_i for each case (which correlate with case typicality)[11]:

$$\alpha_i = p_i^Y - max_{j \neq Y}(p_i^j) \tag{1}$$

In this representation, α_i is the conformity score for the ith case, p_i^Y is the probability that the case is classified correctly, Y is the correct case classification, and $max_{j \neq Y}\ (p_i^j)$ is the maximum probability in the remaining label set, excluding the correct label. We then have a set of conformity scores for these cases, where positive conformity scores represent more typical cases, whereas negative scores represent more atypical cases. After calculating these calibration conformity values, the testing set is run through the classifier to find the predicted labels for the testing cases. With these predicted labels, we can compute the testing conformity of each case using equation 2 below. Testing cases are not associated with a true label, and therefore we must calculate a conformity score for each possible label, defined as a_i^k in equation 2[11]:

$$a_i^k = p_i^k - max_{j \neq ck}(p_i^j) \tag{2}$$

where a_i^k is the conformity score for class k in the ith case, p_i^k is the probability of class label k, and $max_{j \neq ck}(p_i^j)$ is the maximum probability in the remaining label set, excluding label k.

Utilizing the calibration and testing conformity scores, we can calculate the p-values of the testing cases. This allows to transform case conformity into our proposed measures of CAD reliability, confidence, and credibility. To calculate the p-values, we compare each of the testing conformity scores for a case to the set of calibration conformity scores. The p-value[11] is defined in equation 3 as $P_{ik,}$ for class k of the ith case. P_{ik} is shown to equal a_j, the number of calibration conformity scores that are less than or equal to the case conformity score $a_{i\,,}^k$ over the number of calibration conformity scores, $l + 1$. This produces a vector of p-values for a case, which can be used to compute confidence (Cf_i) as one minus the second highest p-value where P_{ij} is the vector of p-values, and credibility (Cr_j) as the maximum p-value in P_{ij}, as defined by Johansson et.al.[11] in equations 4 and 5, respectively.

$$P_{ik} = \frac{\left|\left\{j=1..l\ \&ik\!:\!a_j \leq a_i^k\right\}\right|}{l+1} \tag{3}$$

$$Cf_i = 1 - secondMax_{j=1\dots k}\,(P_{ij}) \tag{4}$$
$$Cr_j = max_{j=1\dots k}\,(P_{ij}) \tag{5}$$

3.4 Encoding the distribution of the label sets

During our testing of our BDT classifier, we observed that many if not most of the nodules with mode ratings of 2 or 4 ended up classified as a 1 or 5 respectively. This led us to the hypothesis that separating these two groups of labels may be generating more noise than would be desirable when training our classifier, and we began searching for ways to combine the label set[23]. The most common approach in the LIDC dataset is to simply divide the labels into high and low rating probabilities, and add the probability of a 3 to one set or the other. One new approach, which we will call *the 2-unweighted-label approach*, more accurately defines a 3 as uncertainty between a high (malignant) or low (benign) rating, and awards half the total uncertain probability to the high and low sets. The values of the probabilities in this 2-unweighted-label approach are given by equations 6 and 7.

$$P(Benign) = P(1) + P(2) + \frac{1}{2}P(3) \tag{6}$$
$$P(Malignant) = P(5) + P(4) + \frac{1}{2}P(3) \tag{7}$$

However, these approaches do not necessarily take advantage of all the information contained in the original 5-rating scale. Instead of this conventional approach, we propose a more *distribution-aware weighting approach* that retains some of the advantages of the 5-rating scale. More specifically, to convert a 5-label probability distribution into a 2-weighted-label distribution, we first look at the meaning of the labels themselves. Since a label of 1 represents an almost certainly benign nodule and a 5 represents an almost certainly malignant nodule, these probabilities do not need to be redistributed. A label of 3 represents complete uncertainty, so in this case, we can split this probability equally between the benign and malignant labels. A label of a 2 or 4 is more likely benign or malignant respectively, but still contains uncertainty. In these cases, we split these probabilities in a weighted manner, giving ¾ to the more probable label, and ¼ to the less probable label as described in equations 8 and 9. For example, a rating distribution of [2, 3, 4, 4], with a BBA of [0 .25 .25 .5 0] will receive 2-weighted-label BBA of [.4375 .5625]. These weighted labels are then used as the new actual labels for constructing the BDTs that will predict probabilities of benign or malignant nodules rather than probabilities of each possible rating.

$$P(Benign) = P(1) + \frac{3}{4}P(2) + \frac{1}{2}P(3) + \frac{1}{4}P(4) \tag{8}$$
$$P(Malignant) = P(5) + \frac{3}{4}P(4) + \frac{1}{2}P(3) + \frac{1}{4}P(2) \tag{9}$$

3.5 Selective iterative classification

In addition to reducing the label set, in a similar effort to improve the BDT's ability to predict malignancy, we implemented the same BDT using selective iterative classification (SIC). To accomplish this, our SIC implementation builds four separate models, each trained on an increasing number of ratings from the training data. When new cases are being classified, they are first run through the model built with one rating. If this model produces a satisfactory prediction, this prediction is final, and no further models are used for classification. If this prediction is unsatisfactory, the case is classified by the next model, and the process repeats. If none of the models produces a satisfactory prediction, then some aggregate prediction of the four models is taken[23,24]. Each BDT classifier was trained and tested using k-fold cross validation[25].

In determining whether a prediction is satisfactory, our implementation uses a technique comprised of three component methods. The first is a maximum probability method; this method looks for a class probability in the predicted distribution > 0.8, and uses this prediction if such a probability is found. If no stopping condition is met, then the classifier moves on to the second method. The second classification method looks for confidence above a certain threshold when determining whether a prediction should be used. If the case is classified with more than a set threshold of confidence by that model, the probability distribution from this model is considered as the result. If it is not, then the same case is run through the next classifier. If none of the models classifies with a confidence over the threshold, then we average the predicted distributions across all four trees to create an aggregate BBA. Threshold values for all three component methods were experimentally determined.

4. RESULTS

The original BDT with conformal prediction and the BDT with SIC were both tested on a balanced set of 850 cases with k-fold cross validation, using 5-label, 2-unweighted-label, and 2-weighted-label distributions. The optimal number of folds for k-fold cross-validation varied between 4 and 6 depending on the BDT implementation and the dataset used. Table 1 summarizes our testing accuracy results where accuracy for the probabilistic classifiers was defined by taking the maximum probability label from the actual and predicted label distributions.

Table 1. Testing accuracy for all tested label sets with either a simple BDT classifier or a BDT classifier with SIC

Accuracy	**Label Distribution Type**		
	5-Label	***2-Label Unweighted***	***2-Label Weighted***
BDT	36.6%	68.9%	76.5%
BDT SIC	45.1%	78.9%	84.0%

The BDT model with a 5-label distribution achieved 36.6% testing accuracy with optimal settings for number of cases per parent n_p equal to 24 and number of cases per children n_c equal to 12. For the same settings, a BDT SIC model achieved 45.1% testing accuracy. On the best fold with a 2-label distribution, our BDT model achieved 74.1% testing accuracy for unweighted labels and 76.5% testing accuracy for weighted labels with the optimal settings of $n_p = 28$ and $nc = 14$. The BDT SIC model achieved 78.9% testing accuracy for unweighted labels and 84.0% testing accuracy for weighted labels with the optimal settings of $n_p = 28$ and $nc = 14$. The conformal prediction thresholds for these results were cf = 0.95 and pr = 0.7; *cf* represents the confidence threshold for SIC models to halt iterative classification, and *pr* represents the probability threshold of any element in a predicted distribution for SIC models to halt iterative classification. As described in the methodology section, if the *pr* threshold is met, the second *cf* threshold is unnecessary, because it already met the first threshold.

To evaluate the proposed approach by taking advantage of the probabilistic distributions of the predicted and target labels, we also calculated the Root Mean Square Error (RMSE) between the actual and predicted distributions. As shown in Figure 1, while the SIC approach had an increase in accuracy for 5-label distributions, it had a slight decrease in mean and standard deviation of the RMSE distribution and a slight increase in kurtosis and skew of the RMSE distribution.

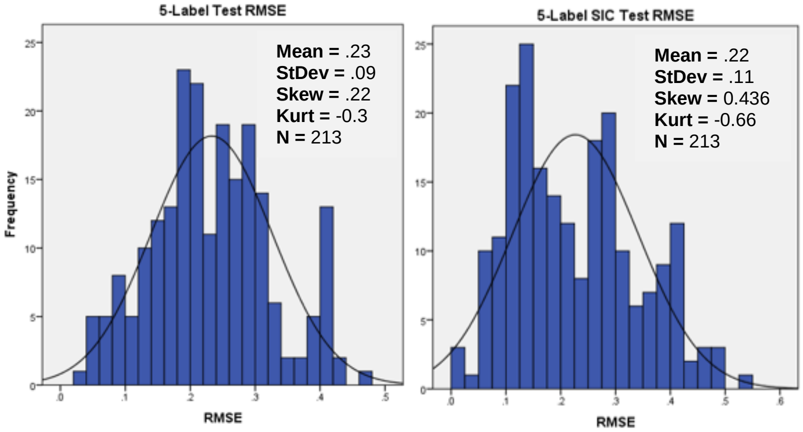

Figure 1. Distribution of RMSE between actual and predicted test labels for the original BDT model (top) and the BDT SIC model (bottom) for the 5-label distributions.

The RMSE results for the unweighted and weighted 2-label models are shown in Figure 2. The weighted models have higher kurtosis values than their counterparts, as well as lower mean RMSE values. This suggests that the weighted model is more consistent in its correct predictions. The SIC approach improves accuracy and RMSE distributions for testing sets on both weighted and unweighted labels. The weighted SIC model produced the lowest mean RMSE, maximum RMSE, and standard deviation RMSE, as well as the highest accuracy of any 2-label model.

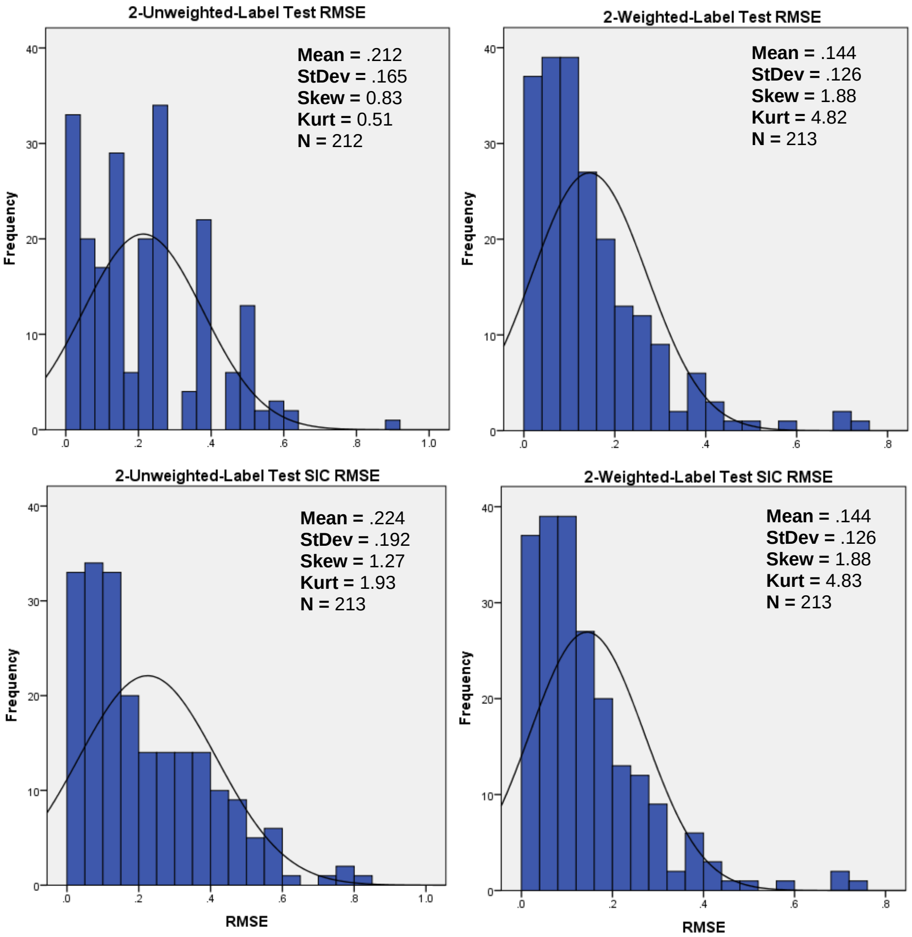

Figure 2. Distribution of RMSE between actual and predicted testing labels for: the original BDT model (top) and the BDT SIC model (bottom) with the 2-unweighted-label distribution (left) and the 2-weighted-label distribution (right).

5. CONCLUSIONS

Our proposed approach to combine selective iterative classification with conformal prediction produced better accuracy results than when using classification approaches based on only belief decision trees. Furthermore, taking the advantage of the meaning of the label data, we showed that the results can be further improved by associating weights to the labels, with a higher weight for the most likely label and a lower weight for the least likely label. By incorporating conformal prediction into the decision-making process of the iterative probabilistic classifiers, we were not only improving accuracy, but we also provided measures of confidence and credibility for the CAD predictions. In the long run, this additional information about the CAD output can increase the trust in the CAD outcome, and therefore, impact the large-scale adoption of these systems in clinical use.

ACKNOWLEDGEMENTS

This work has been supported by the NSF Research Experiences for Undergraduates (REU) site in Medical Informatics grant #1359459.

REFERENCES

[1] Karssemeijer N, Risso G, Catarzi S, et al: Computer-aided detection versus independent double reading of masses on mammograms. Radiology 227:192–200, 2003

[2] Mower WR: Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 33(1):85–91, 1999

[3] Jarvik JG, Deyo RA: Moderate versus mediocre: the reliability of spine MR data interpretations. Radiology 250(1):15–17, 2009

[4] MacMahon H, Engelmann R, Behlen F, Hoffmann K, Ishida T, Roe C, Metz C, Doi K: Computer-aided diagnosis of pulmonary nodules: Results of a large-scale observer test. Radiology 13:723– 726, 1999

[5] 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 higher resolution CT: evaluation with receiver operating characteristic analysis. Am J Roentgenol 178(3):657–663, 2002

[6] Li F, Aoyama M, Shiraishi J, et al: Radiologists' performance for differentiating benign from malignant lung nodules on high resolution CT using computer estimated likelihood of malignancy. Am J Roentgenol 183:1209–1215, 2004

[7] Marten K, Grillhösl A, Seyfarth T, Obenauer S, Rummeny EJ, Engelke C: Computer-assisted detection of pulmonary nodules: evaluation of diagnostic performance using an expert knowledge based detection system with variable reconstruction slice thickness settings. Eur Radiol 15:203–212, 2005

[8] Fletcher JW, Kymes SM, Gould M, Alazraki N, Coleman RE, Lowe VJ, et al: A comparison of the diagnostic accuracy of 18FFDG PET and CT in the characterization of solitary pulmonary nodules. J Nucl Med 49:179–185, 2008

[9] Bankier AA, Levin D, Halpern EF, Kressel HY: Consensus interpretation in imaging research: is there a better way? Radiology 257:14–17, 2010

[10] Zinovev, D., Duo, Y., Raicu, D. S., Furst, J., Armato, S. G., "Consensus Versus Disagreement in Imaging Research: a Case Study Using the LIDC Database," Journal of Digital Imaging 25(3), 423–436 (2011).

[11] Johansson, U., Bostrom, H., & Lofstrom, T. (2013). Conformal Prediction Using Decision Trees. *2013 IEEE 13th International Conference on Data Mining*.

[12] Armato, S. G., *et al*. (2004). Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community. *Radiology, 232*(3), 739-748.

[13] Drukker, K., Sennett, C., & Giger, M. (2009). Automated Method for Improving System Performance of Computer-Aided Diagnosis in Breast Ultrasound. *IEEE Transactions on Medical Imaging IEEE Trans. Med. Imaging, 28*(1), 122-128.

[14] Jagdale, S., Kolekara, M. H., & Khot, U. P. (2015). Smart Sensing Using Bayesian Network for Computer Aided Diagnostic Systems. *Procedia Computer Science, 45*, 762-769.

[15] Marzieh Kohandani Tafreshi, Virendra Joshi, Alexander Meining, Charles Lightdale, Marc Giovannini, et al.. Smart Atlas for Supporting the Interpretation of probe-based Confocal Laser Endomicroscopy (pCLE) of Biliary Strictures: First Classification Results of a Computer-Aided Diagnosis Software based on Image Recognition. Digestive Disease Week (DDW 2014), May 2014, Chicago, United States. 2014.

[16] Papadopoulos, H., Gammerman, A., Vovk, V. (2009). Reliable diagnosis of acute abdominal pain with conformal prediction. *Engineering Intelligent Systems.*

[17] Riely A., Sablan K., Fang X., Furst J.D., Raicu D.S. "Reducing Annotation Cost and Uncertainty in Computer-Aided Diagnosis through Selective Iterative Classification", SPIE Medical Imaging 2015

[18] Zinovev, D., Feigenbaum, J., Furst, J., & Raicu, D. (2011). Probabilistic lung nodule classification with belief decision trees. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

[19] Carrazza M., Kennedy B., Furst J.D., Rasin A., Raicu D.S. "Investigating the effects of majority voting on CAD systems: an LIDC case study", SPIE Medical Imaging. San Diego, California, February 27- March 3, 2016

[20] Ochs, R., Kim, H. J., Angel, E., Panknin, C., Mcnitt-Gray, M., & Brown, M. (2007). Forming a reference standard from LIDC data: Impact of reader agreement on reported CAD performance. *Medical Imaging 2007: Computer-Aided Diagnosis*.

[21] D. Zinovev, D. Raicu, J. Furst, S. G. Armato III, "Predicting radiological panel opinions using a panel of machine learning classifiers," Algorithms Journal, vol. 2, 2009, pp. 1473-1502.

[22] Elouedi, Z., Mellouli, K., & Smets, P. (2001). Belief decision trees: Theoretical foundations. *International Journal of Approximate Reasoning, 28*(2-3), 91-124.

[23] Whitehill, J., Wu, T., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Neural Information Processing Systems (NIPS),* (22).

[24] Ji, M., Han, J., & Danilevsky, M. (2011). Ranking-based classification of heterogeneous information networks. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*.

[25] Fushiki, T. (2009). Estimation of prediction error by using K-fold cross-validation. Statistics and Computing Stat Comput, 21(2), 137-146.