**Building confidence and credibility into CAD with belief decision trees**

Rachael N. Affenit2, Erik R. Barns1, Jacob D. Furst1, Alexander Rasin1, Daniela S. Raicu1  
1 College of Computing and Digital Media, DePaul University, Chicago, IL, US  
2 College of Science, Illinois Institute of Technology, Chicago, IL, US

**ABSTRACT**

To improve on existing CAD systems, we first examined Belief Decision Trees to provide a vector of five predicted probabilistic labels of malignancy for each case in a balanced version of the Lung Image Database Consortium (LIDC) dataset. We then applied conformal prediction to these results in order to analyze the reliability of the predictions. These probabilistic predictions given with levels of confidence and credibility can build a smarter CAD system that provides more contextual information to the radiologist. The BDT with conformal prediction was validated for a balanced set of 850 cases with k-fold cross validation. On the best fold, this implementation achieved 45% testing accuracy when predicting a 5-label distribution (np = 24, nc = 12, dmax = 25, k = 6).

Our model achieved 85.4% testing accuracy when predicting a 2-label distribution (np = 16, nc = 8, dmax = 12, and k = 4).

This method was then applied to a modified dataset, containing only two aggregated probabilistic labels representing the probability of a benign or malignant case. This BDT was modified to include selective iterative classification, which incorporates confidence from conformal prediction into the classification algorithm as a form of self-evaluation, and run on both 5-label and 2-label distributions. On the best fold, the SIC classifier achieved 37.0% testing accuracy for a 5-label distribution, and 82.5% testing accuracy for a 2-label distribution.

**Keywords:** Computer-aided diagnosis, LIDC, belief decision tree, conformal prediction, reliability, confidence, credibility

1. **INTRODUCTION**

Computer-aided diagnosis (CAD) systems provide radiologists with supplemental diagnostic information for use when analyzing patient CT or X-Ray images. These systems reduce the work required to assess an image by quantitatively analyzing images and predicting qualitative characteristics of the case. However, CAD systems currently have a relatively low adoption rate in clinical setting, in part because radiologists do not necessarily trust the results due to the lack of supporting contextual information. Without giving clinicians a window into the CAD system’s reasoning, or estimates of how well it can predict a particular case, radiologists may come to distrust its results as they observe more misclassified cases [1]. A more comprehensive statistical and contextual output for each case may help build their trust in these types of systems, and give clinicians a better understanding of how well a prediction fits each case. By providing probabilistic labels for potential diagnoses, along with self-evaluation features like prediction reliability.

We modified a BDT with conformal prediction that we had previously constructed [CITE???] to create a classifier that incorporates Selective Iterative Classification (SIC) based on confidence of a predicted distribution. These two classifiers were first applied to the original data with a 5-label probability distribution, then applied to a modified 2-label dataset, containing only two aggregated probabilistic labels representing the probability of a benign or malignant case.

1. **RELATED WORK**

**2.1 Smart CAD**

Various smart CAD systems have been developed over the years. Drukker, et. al. [1] developed a CAD system for breast ultrasound in 2009, which examined the difference between calculated nodule boundaries and radiologist marked nodule boundaries to determine its confidence in a diagnosis. They also examined different uses for this confidence measure, including as output to the radiologist, or as a self-evaluation measure to auto assess the performance of a specific classifier when using multiple classifiers in the system. Jagdale, et. al. [2] developed a CAD system for mammography, which used a Bayesian network classifier to distinguish tumor cells from healthy tissue. Marzieh et. al. [3] developed a system called Smart Atlas to identify biliary structures from confocal laser endomicroscopy (pCLE) video. They used measures of specificity and positive predicted value (PPV) to provide contextual information to the medical professional.

**2.2 Lung Image Database Consortium (LIDC)**

Iii, et. al. [4] developed the LIDC in 2004, surveying between 1 and 4 radiologists to determine eight semantic characteristics of lung nodules: malignancy, internal structure, calcification, margin, lobulation, sphericity, spiculation, and texture. In this paper, we examine the subset of 809 cases that were rated by 4 radiologists to classify the cases in terms of their malignancy rating. Ochs, et. al. [5] studied the impact of rater agreement on classifier performance in LIDC dataset.

**2.3 Belief decision trees (BDT)**

Elouedi, et. al. [6] outlined the mathematical theory of Belief Functions, and how to incorporate them into the framework of a typical decision tree in order to create a Belief Decision Tree (BDT). These BDTs are designed to provide probabilistic classifications on uncertain data. Zinonvev, et. al. [7] described how this type of decision tree can be used to make probabilistic predictions for cases in the LIDC dataset.

**2.4 Conformal prediction (CP)**

Johannson, et. al. [10] described a mathematical method of integrating conformal prediction with decision trees, based on its implementation with several other machine learning algorithms. Conformal prediction has also been used by Harris et. al. [11] in conjunction with CAD for acute abdominal pain.

**2.5 Selective iterative classification (SIC)**

Ji, et. al. implemented an algorithm called RankClass, an iterative classifier that is integrated with the rankings used for each iteration [8] Whitehill, et. al. [9] investigated the wide range of levels of expertise that could be present in the raters, how difficult a specific image is to classify, the combination of multiple labels to improve classification.

**2.6 K-Fold cross validation**

Fushiki, T. [12] described how to use K-Fold Cross Validation to get a better sense of the performance of a classifier on a small dataset.

1. **METHODOLOGY**

The lack of consensus between semantic ratings in the LIDC dataset can introduce unwanted bias when classifying new nodules as benign or malignant. In order to deal with this uncertainty, we have implemented an algorithm based on a probabilistic classifier called Belief Decision Tree (BDT). It is an adaption of a decision tree (DT) classifier that uses belief function theory to better handle uncertainty. On top of our BDT, we have implemented conformal prediction (CP), to calculate measures of reliability for each predicted probability distribution. We also examined the possibility of reducing the label set from 5 probabilistic labels to 2, as treating the 1-5 scale like separate ratings did not always yield the most accurate of results; ratings of 2 and 4 were frequently misclassified as 1 and 5 respectively, as their image features should be relatively similar.

**3.1 LIDC dataset**

The LIDC dataset [4] contains between one and four radiologist ratings of a nodule’s malignancy for each case, on a scale of 1 (benign) to 5 (malignant), where 3 represents uncertainty. Unfortunately, in this setting a radiologist rating is not akin to a ground truth in machine learning, and the four radiologists agree on a consensus label in only 25% of these cases [5]. A set of 64 image features had been previously calculated for each case in this dataset, and threshold values at these features are chosen by the algorithm to determine how the BDT is built. We are working with a subset of this data which includes only those 809 cases where all four radiologists identified and rated a nodule. This subset was then balanced by under sampling the uncertain (label 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.

**3.2 Belief decision trees**

In a similar vein to a decision tree, a BDT classifies an LIDC case by comparing calculated image feature values to the chosen threshold values to determine which path in the tree the case should follow. When a case reaches a leaf node, it can be assigned a Basic Belief Assignment (BBA) associated with this node as a method of classifying that case. This BBA is a set of probabilities for each of the five classification labels, and represents the average BBA of all cases in the training set that reached this node [6]. The training case BBA’s were created using the radiologist ratings from the dataset; for example, a rating distribution of [2, 3, 4, 4] would yield the BBA [0 .25 .25 .5 0] for five label distribution. Typically the process of calculating these probabilities is much more involved for a belief decision tree, but the LIDC dataset has a few special qualities that allow us to use this method. Every radiologist can only choose one malignancy rating for each case, which allows us to eliminate the possibility of having two or more “true” labels. The dataset also has no representation of pure uncertainty (a rating of a 3 indicates balanced probabilities of malignant or benign labels), resulting in a simplification of this calculation to a probability distribution [7].

The biggest difference between a decision tree and a BDT occurs during tree construction. When deciding whether and how a node should split, a BDT calculates the pignistic probabilities of each class for every case in the dataset (which becomes our BBA), and averages the probabilities of all the cases that reach each node in the tree. The average pignistic probabilities of the parent and child nodes can then be used to calculate the information gain of splitting, using each possible feature and threshold value in the dataset. It then computes the gain ratio, which controls for the size of the child subsets and rewards equally distributed splits, and chooses the feature and threshold that achieved the maximum gain ratio for the split. One can determine whether a node in a BDT is a leaf node if it meets one of four stopping criterion: the maximum information gain of splitting was 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 np and nc parameters), all of the BBA’s at the node are equivalent, or all features have already been used to split [6].

**3.3 Conformal prediction**

On top of our BDT, we implemented Conformal Prediction (CP) in our BDT to produce measures of confidence and credibility for each CAD probability distribution. CP 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 testing 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 for each case (which correlate with case typicality). 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. Johansson et.al. [10] defines calibration conformity as in Eq. 1 below. In this representation, is the conformity score for the ith case, is the probability that the case is classified correctly, and is the maximum probability in the remaining label set, excluding the correct label.

Eq. 1

After these calibration conformity values have been calculated, 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 Eq. 4 below. Testing cases are not associated with a true label, and therefore we must calculate a conformity score for each possible label, defined as by Johansson et.al. [10]. Shown in Eq. 2, is the conformity score for class k in the ith case. is the probability of class label k and is the maximum probability in the remaining label set, excluding label k.

Eq. 2

Utilizing the calibration and testing conformity scores, we can calculate the p-values of the testing cases. This allows us to transform case conformity into our measures of 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. It represents the ratio of conformity scores in the calibration set that are less than or equal to the conformity score of that label to the total number of instances in the calibration set. Using this p-value, we can calculate confidence and credibility for each prediction as both an output for the radiologist to consider, and as a method of choosing the best BBA during iterative classification. The p-value is defined by Johansson et.al. [10] in Eq. 5 as , for class k of the ith case. is shown to equal the number of calibration conformity scores that are less than or equal to the case conformity score , over the number of calibration conformity scores, . This produces a vector of p-values for a case, which can be used to compute confidence () as one minus the second highest p-value where  is the vector of p-values, and credibility () as the maximum p-value in , as defined by Johansson et.al. [10] in Eq. 4 and 5.

Eq. 3

Eq. 4

Eq. 5

**3.4 Intelligently reducing the label set**

During our testing of our BDT classifier, we observed that many if not most of the nodules with mode ratings of 2 or 4 would end 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 reduce the label set [9]. 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 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 2-unweighted-label approach is described by Eq. 7 and 8 below. However, these approaches do not necessarily take advantage of all the information contained in the original 5-rating scale.

Eq. 7  
 Eq. 8

Instead of this more conventional approach, we propose a more intelligently weighted approach that retains some of the advantages of the 5-rating scale while reducing the noise present when training the classifier. In order to convert a 5-label probability distribution into a 2-weighted-label distribution, we first broke down the meaning of the labels themselves. In doing so, we discovered that the five labels of malignancy can be easily reduced to two labels with varying levels of certainty A label of 1 represents an almost certainly benign nodule, whereas a 5 represents an almost certainly malignant nodule, so 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. So, for the example rating distribution of [2, 3, 4, 4], and our original BBA of [0 .25 .25 .5 0], our new 2-weighted-label BBA would be [.4375 .5625]. These converted labels were then used as the “Actual” labels for our tree, which then would predict probabilities of benign or malignant nodules rather than probabilities of each possible rating. This weighted approach is described in Eq. 9 and 10 below.

Eq. 9  
 Eq. 10

**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). The aim of SIC is to control for some of the uncertainty between ratings, given unknown expertise of raters. In order to accomplish this, our SIC implementation builds four separate classifiers, each with an increasing number of radiologist ratings considered in its training set. A case is classified using the mode classification of three different classification methods. This includes a max BBA method that uses the tree classification with a BBA probability of over 0.8 as a stopping condition or the BBA with the highest probability across all four trees. The second classification method uses the confidence level of the tree output to select an optimal classification. If the case is classified with more than 0.90 confidence by that model, the probability distribution from this model is considered the final result. If it is not, then the same case is run through the next classifier. If none of the models classify with a confidence over the threshold, then we BBA classification that produces the highest confidence. Lastly we use a non-selective classification method of averaging the BBA outputs across all four trees. Lastly we average the BBA output across all three iterative classification methods as our final output. [Need to reference 8,9]

**3.6 K-Fold cross validation**

For assessing the performance of our BDT with conformal prediction, we decided to use k-fold cross validation [12]. This technique splits the full dataset into k subsets, and the classifier is run k times on different combinations of sets, with each subset acting as a testing set once. The remaining k – 1 subsets for each combination constitute the training set, and once these sets are chosen, we split the training set into the proper training set and the calibration set by randomly pulling 1/7 of the cases to be used for calibration. We decided to report the accuracy for the best fold, as our small dataset tends to distort the average accuracy over all folds. The optimal number of folds varied between 4 and 6 depending on the BDT implementation and the dataset used. When attempting to use the recommended number of folds of k = 10, our testing set was too small to produce meaningful results.

1. **RESULTS**

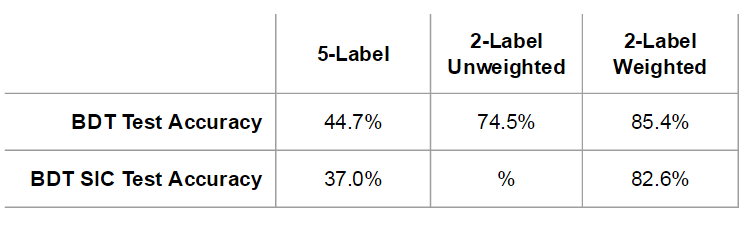
Two models were used to obtain results: the original BDT with conformal prediction, and that BDT with SIC. Both were validated on a balanced set of 850 cases with k-fold cross validation, using 5-label, 2-unweighted-label, and 2-weighted-label distributions.

On the best fold with a 5-label distribution, our BDT model achieved 45% testing accuracy with optimal settings of np = 24, nc = 12, dmax = 25, and k = 6. The BDT SIC model was able to achieve 37% testing accuracy; the optimal settings for the one and two rating trees were np = 16, nc = 8, dmax = 12, k = 5, and cf = 0.6, while the three and four rating trees had np = 20, nc = 10, dmax = 12, k = 5, and cf = 0.6.

On the best fold with a 2-unweighted-label distribution, our BDT model achieved 85.4% testing accuracy with the optimal settings of np = 16, nc = 8, dmax = 12, and k = 4. The BDT SIC model was able to achieve 82.6% testing accuracy; the optimal settings for the one and two rating trees were np = 16, nc = 8, dmax = 12, k = 5, and cf = 0.9, while the three and four rating trees had np = 20, nc = 10, dmax = 12, k = 5, and cf = 0.9.

On the best fold with a 2-weighted-label distribution, our BDT model achieved 85.4% testing accuracy with the optimal settings of np = 16, nc = 8, dmax = 12, and k = 4. The BDT SIC model was able to achieve 82.6% testing accuracy; the optimal settings for the one and two rating trees were np = 16, nc = 8, dmax = 12, k = 5, and cf = 0.9, while the three and four rating trees had np = 20, nc = 10, dmax = 12, k = 5, and cf = 0.9.

**Table 1.** Shows BDT and BDT SIC test accuracy values for 5-label, 2-unweighted-label, and 2-weighted-label distributions



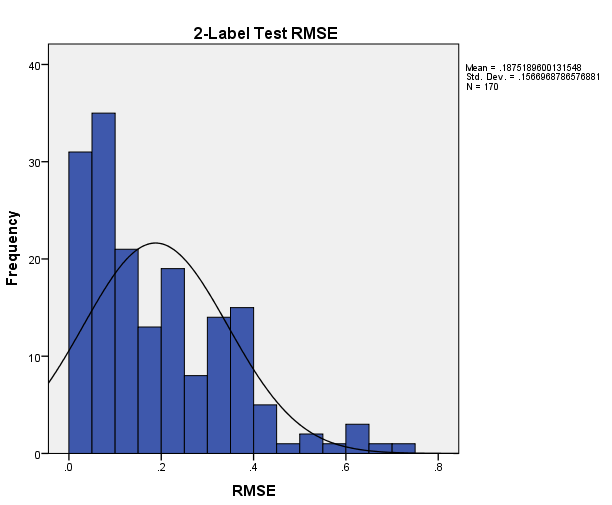
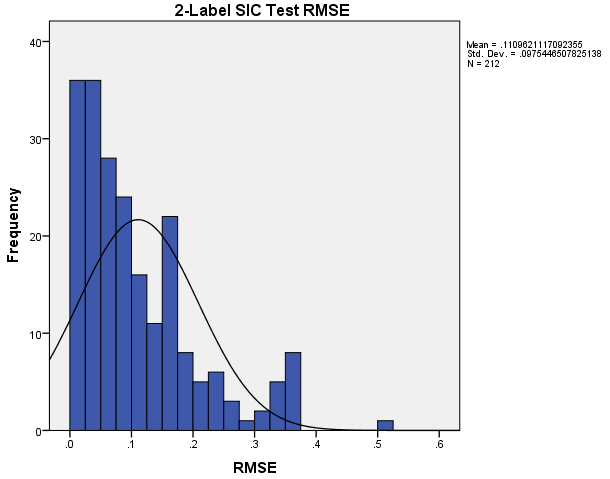
For these settings, np represents the minimum number of cases a parent node must have to split, nc represents the minimum number of cases a child node must have for its parent to split, dmax represents the maximum tree depth, k represents the number of folds used for the validation of the tree, and cf represents the confidence threshold for a predicted distribution above which BDT SIC models halt iterative classification. Accuracy and other deterministic evaluation methods were defined by taking the maximum probability label from the actual and predicted label distributions, and assuming these were the actual and predicted labels respectively. The accuracy values obtained are likely influenced by the small size of our subset, and the necessity of splitting it further into even smaller subsets to use conformal prediction. Accuracy may also be lowered by forcing a consensus out of intentionally uncertain probability distributions as well, in which case accuracy is likely not the best measure of performance.

In order to better compare the actual and predicted distributions without forcing consensus, we calculated the RMSE between the actual and predicted distributions, shown in Fig. 1, which helps us assess the BDT’s ability to predict the correct distribution. RMSE for our sample size of 212/213 testing pairs is significant with p < 0.05 at any correlation above \_\_\_\_.

When we compare the RMSE distributions between the BDT SIC model (left) and the original BDT model (right) performance in Fig. 1, it has a lower mean and standard deviation with a similar with the inclusion of SIC. The BDT SIC model also achieved a 0.93 AUC compared to the original BDT’s 0.00 AUC (ROC curves in Fig. 2).

[Compare all three threshold methods of SIC]

[How does random selection of each label for SIC affect results?]

****

**Mean =** .188  
**StDev =** .157

Skew = 1.02

Kurt = 0.70

**N =** 213

**Mean =** .111

**StDev =** .098

**Skew =** 1.35

**Kurt =** 1.64  
**N =** 212

**Figure 1.** Distribution of RMSE between actual and predicted test labels for the BDT SIC model (left) and the original BDT model (right) for the 2-weighted-label distributions

[Find ROC AUC in SPSS (predicted, actual, 1 for “malignant” probability)]

**Figure 2.** ROC curves for the BDT SIC model (left) and the original BDT model (right) for the 2-weighted-label distributions

1. **DISCUSSION**

By incorporating conformal prediction into the decision-making process of an iterative probabilistic classifier, we are proposing a CAD system that will provide more informative probabilistic predictions for new cases, which would include measures of confidence and credibility for those predictions. In so doing, we aim to generate results that give clinicians a better idea of the context surrounding the predictions, and build their trust in the viability CAD tools for clinical use.

**REFERENCES**

[1] 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.

[2] 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.

[3] 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.

[4] Iii, S. G., Mclennan, G., Mcnitt-Gray, M. F., Meyer, C. R., Yankelevitz, D., Aberle, D. R., . . . Clarke, L. P. (2004). Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community1. *Radiology,* *232*(3), 739-748.

[5] 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*.

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

[7] 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*.

[8] 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*.

[9] 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).

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

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

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