**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). [CONCLUSIONS]

**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.

In order to dispel some of the doubt in the abilities of CAD systems, we investigate strategies to produce better prediction performance, as well as measures of prediction reliability. We have examined Belief Decision Trees (BDT) to provide a vector of predicted probabilistic labels of malignancy for each case in a balanced version of the Lung Image Database Consortium (LIDC) dataset, and applied conformal prediction (CP) to these results in order to analyze the reliability of the classifier’s predictions. Probabilistic predictions given with levels of confidence and credibility from conformal prediction can be used to build a smarter CAD system that provides more contextual information to the radiologist.

1. **RELATED WORK**

**2.1 Smart CAD with Performance Evaluation**

Various “Smart CAD” systems with built-in performance evaluation features have been developed in recent years for different modalities in medical imaging. 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. This 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 this ROI. Marzieh et. al. [3] 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.

**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, finding that cases with higher initial reader agreement on a rating for a semantic characteristic were also classified with a higher sensitivity by the CAD system.

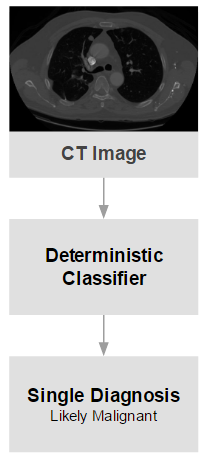
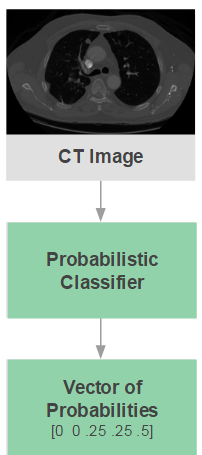
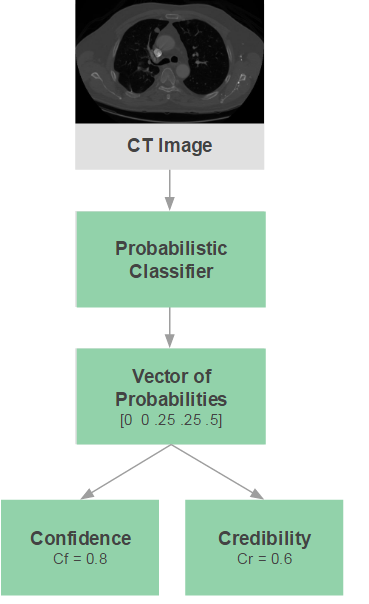
**2.3 Belief decision trees (BDT)**

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. The single BDT from this paper predicted a distribution of malignancy ratings with 48.8% testing accuracy on a dataset of 914 nodules (10% validation, 59% training, 31% testing).

**2.4 Conformal prediction (CP)**

Conformal prediction is a method of determining the reliability of a classifier’s prediction. 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. 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 a CAD system for acute abdominal pain. This system relied on neural networks as the base classifier, and achieved 75.7% correct diagnoses.

1. **METHODOLOGY**

**Figure 1.** Typical CAD system input and output with a deterministic classifier (left), CAD system input and output with a probabilistic classifier (middle), and CAD system input and output with both a probabilistic classifier and conformal prediction (right).

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.

**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

**3.6 K-Fold cross validation**

For assessing the performance of our BDT with conformal prediction, we decided to use k-fold cross validation [12]. For each fold in the validation process, we have a unique BDT, training set, testing set, and calibration set. This technique splits the full balanced dataset into k subsets, from which we use k – 1 subsets as the training set and the remaining subset as the testing set. 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. The classifier is run k times on different combinations of these subsets, with each subset acting as a testing set once. 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.

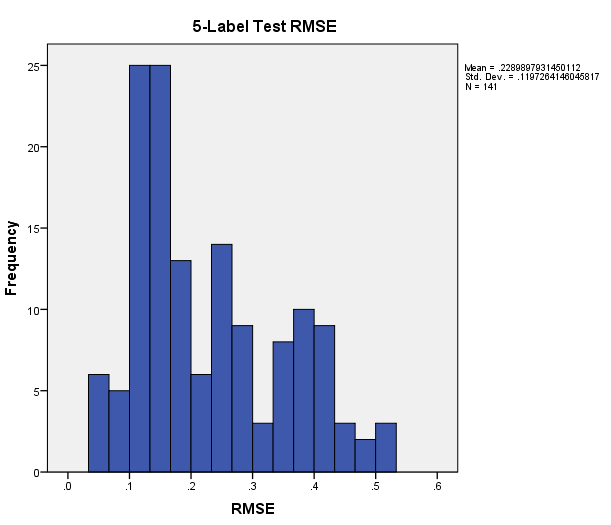
1. **RESULTS**

The BDT was implemented with conformal prediction, and validated on a balanced set of 850 cases with k-fold cross validation. On the best fold, this implementation achieved 45% testing accuracy, with its optimal settings were found to be np = 24, nc = 12, dmax = 25, and k = 6.

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, and k represents the number of folds used for the validation of the tree. Accuracy was defined by taking the maximum probability label from the actual and predicted label distributions to be 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 three even smaller subsets. 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.

To compare the actual and predicted distributions without forcing consensus, we calculated the RMSE between the actual and predicted distributions, shown in Fig. 2, which helps us assess the BDT’s ability to predict the correct distribution. RMSE for our sample size of 141 testing pairs is significant with p < 0.05 anywhere above \_\_\_\_. This allows us to conclude that \_\_\_\_% of test cases have a statistically significant correlation between actual and predicted distributions.

[Include # typical versus # atypical cases in the full dataset]

****[Confidence by label] [Typicality]

**Mean =** .228  
**StDev =** .119

**N =** 141

**Figure 2.** Distribution of RMSE between actual and predicted test labels for our classifier (left), confidence distributions for each of the five labels in our test set (middle), and typicality of all cases in the balanced dataset (right).

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.