Label Set Reduction for Computer-aided Diagnosis with Uncertainty

Erik R. Barns, Jacob D. Furst, Daniela S. Raicu

College of Computing and Digital Media

DePaul University

Chicago, Illinois

erik.barns909@gmail.com

Rachael N. Affenit

Department of Computer Science

Illinois Institute of Technology

Chicago, Illinois

raffenit@gmail.com

*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 in a fashion that takes into account the degrees of malignancy on a scale from 1 to 5 with 1 being “most likely benign” and 5 being “most likely malignant”. Using the NIH/NCI Lung Image Database Consortium (LIDC) data, we show that the proposed 2-label-weighted approach significantly outperforms the original 5-label and 2-label-unweighted classification problems by X and Y, respectively, for accuracy and Z and Q, respectively, for the average root mean squared error (RMSE) between the predicted and actual distributions. Furthermore, to take advantage of the multiple experts’ interpretations, we also incorporate a selective iterative classification approach (SIC) based on conformal prediction. The SIC approach improves overall accuracy for weighted and unweighted models by M and N, as well as improves the average RMSE of the 2-label weighted model by P and R.

Keywords—Computer-aided diagnosis; LIDC; belief decision tree; confidence; selective iterative classification; label set reduction

# Introduction

Early detection and diagnosis is critical in cases of potential lung cancer. Misdiagnosis of a malignant nodule, no matter how small, can significantly impact a patient’s chances of survival. Smaller malignant nodules are more frequently missed on CT radiographs, simply due to the difficulty of visually detecting them in the image. In these types of cases, image processing techniques can provide valuable contextual information. Computer-aided diagnosis (CAD) systems can analyze these types of images for specific calculable image features, and provide radiologists with supplemental information about an image when analyzing patient radiographs. These systems can reduce the work required to assess an image, and help the radiologist detect smaller nodules by quantitatively analyzing image features from a radiograph and highlighting any concerning features.

Most of these CAD systems are trained on radiologists’ opinions of a case rather than a final diagnosis, so there can be a significant amount of variability in this data. Some systems tackle this problem by outputting measures of prediction reliability for use by the radiologist.

By incorporating these reliability measures into the classification process itself, we hypothesize that we can build a more robust classifier that can perform better with noisy training data. These measures can be incorporated into another technique for handling noise in a classifier’s training data: selective iterative classification. A case would be classified by a model trained on one rating, then if the confidence of this prediction is above a certain threshold, this prediction would be assigned to the case. If not, then the case would be classified by a model trained on an additional rating- this process continues until no more ratings remain. An additional technique we wish to investigate is the possibility of reducing the label set. For the LIDC dataset we are using, the radiologist ratings of malignancy can be converted into a 5-label probability distribution, where 1 represents a benign case and 5 represents a malignant case. Instead of having this distribution with varying levels of uncertainty, we intend to calculate the probability of a case being benign or malignant based on the distribution of uncertainty in the original 5-label distribution.

Our classifier is based on a belief decision tree (BDT) with conformal prediction. We tested this classifier on its own, as well as a version that incorporates selective iterative classification (SIC), on a balanced set of 850 cases with k-fold cross validation. This testing process was repeated for a 5-label distribution, a simple 2-unweighted-label distribution, and our proposed 2-weighted-label distribution.

# Related Work

## CAD for Lung Image Database Consortium (LIDC)

Armato, 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. Rater variability is not the only type of uncertainty present in the LIDC dataset; in fact, there is uncertainty on multiple levels, including disagreement between radiologists, ambiguity of labels, and the partially uncertain and entirely uncertain labels themselves. Many CAD systems choose to handle uncertainty in labels by finding a consensus using majority voting, but this technique results in the loss of a significant amount of information when producing a consensus.

## Belief decision trees (BDT)

Rather than attempting to produce a consensus label, belief decision trees simply produce a vector of possibilities for each label. 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).

## Selective iterative classification (SIC)

Since the LIDC dataset contains 4 separate radiologist ratings of malignancy, we can examine the benefits of selective iterative classification. SIC is used to make sense of noisy training data when there is no real ground truth to train on; the effect of this would be similar to majority voting, except the classifiers contribute to a final probability distribution rather than a final label. It involves building r separate models, each trained on a different number of ratings from the training data, where r is the maximum number of ratings 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. Ji, et. al. [8] implemented an algorithm called RankClass for heterogeneous information networks, an iterative classifier that is integrated with the rankings used for each iteration. When classifying a subset of the DBLP dataset in four specific research areas, this algorithm was able to produce an average accuracy of 88% on author objects, 83% on paper objects, and 90% on conference objects. 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. Their paper then proposes a probabilistic classifier to label a series of images, which outperforms the majority vote heuristic and is able to better handle noise in its training data.

## 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. The aim of these systems is to provide extra contextual information to the radiologist in order to better inform the final diagnosis. 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.

## Conformal Prediction (CP)

Conformal prediction is a method of determining the reliability of a classifier’s prediction, and can be used as a type of performance evaluation for a CAD system. 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.

## Our Contribution

We build on all of this previous work by investigating the merits of performing selective iterative classification on a BDT using conformal prediction, which provides a self-correcting feature in the CAD system. In parallel, we begin to investigate methods of reducing the label set and how they affect the performance of the model. Reducing a label set is known to increase classification accuracy, and the aim is to achieve this while retaining some of the useful information that could be gained from a distribution of five malignancy labels in the LIDC dataset. Both of these techniques would be applicable to any Smart CAD system using probabilistic classifiers, and can significantly improve the adoption of these systems by clinicians and hospitals [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.

## 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.

## 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].

## 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 *αi* in (1) below. In this representation, is the conformity score for the ith case, *piY* is the probability that the case is classified correctly, and *maxj≠Y* (*pij)* is the maximum probability in the remaining label set, excluding the correct label.

*αi**piY – maxj≠Y**pij)* 

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 (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 *αik* by Johansson et.al. [10]. Shown in (2), *αik* is the conformity score for class k in the ith case. *pik* is the probability of class label k and *maxj≠ck* (*pij)* is the maximum probability in the remaining label set, excluding label k.

*αik**pik – maxj≠ck**pij)* 

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 (3) as *Pik*, for class k of the ith case. *Pik*  is shown to equal *aj,* the number of calibration conformity scores that are less than or equal to the case conformity score *aik*, 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, *Cfi*, as one minus the second highest p-value where *Pij* is the vector of p-values, and credibility, *Crj*, as the maximum p-value in *Pij*, as defined by Johansson et.al. [10] in (4) and (5).

*Pik**(j = 1…l & ik : aj ≤ aik)|*   
*l + 1*

*Cfi* *secondMaxj = 1…k**Pij)* 

*Crj* *maxj = 1…k**Pij)* 

## 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 (6) and (7) below. However, these approaches do not necessarily take advantage of all the information contained in the original 5-rating scale.

*P(Benign) = P(1) + P(2) + ½P(3)* 

*P(Malignant) = P(5) + P(4) + ½P(3)* 

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 (8) and (9) below.

*P(Benign) = P(1) + ¾P(2) + ½P(3) + ¼P(4)* 

*P(Malignant) = P(5) + ¾P(4) + ½P(3) + ¼P(2)* 

## 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 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 [8, 9].

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 probability threshold is not met, the classifier moves on to the second method. The second classification method iterates through the four models and uses the output from the model that produces a confidence level above a certain threshold. If it is not, then the same case is run through the next classifier. If none of the models classify with a probability or confidence over the set threshold, then we result in using the output from a third method. The third classification component uses the predicted distributions across all four trees to create an aggregate BBA. Threshold values for both component methods were experimentally determined by the average maximum probability and confidence of the cases output that were classified correctly by a BDT that uses four radiologist ratings.

## 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 in 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.

# Results

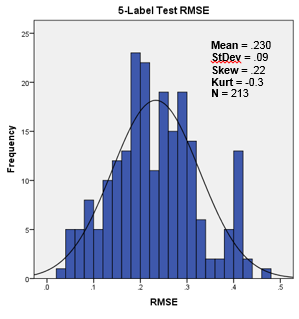
Two models were used to obtain results: the original BDT with just conformal prediction, and the BDT with SIC. Both were validated on a balanced set of 850 cases with k-fold cross validation, using 5-label, 2-unweighted-label, and 2-weighted-label distributions. Each of the threshold methods for SIC was examined on the 2-weighted-label distribution. Test accuracy values on the best fold for all models are listed in Table 1.

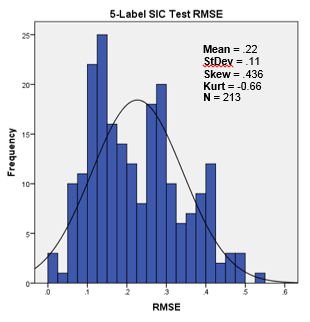
1. Testing Accuracy for All Tested Models

| 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% |

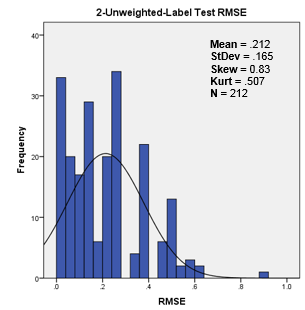
On the best fold with a 5-label distribution, our BDT model achieved 36.6% testing accuracy with optimal settings of *np = 24, nc = 12, dmax = 12, k = 5.* The BDT SIC model was able to achieve 45.1% testing accuracy with optimal settings of *np = 24, nc = 12, dmax = 12, k = 4, cf = 0.95, pr = 0.5*. On 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 *np = 28, nc = 14, dmax = 12, k = 4*. The BDT SIC model was able to achieve 78.9% testing accuracy for unweighted labels and 84.0% testing accuracy for weighted labels with the optimal settings of *np = 28, nc = 14, dmax = 12, k = 4, cf = 0.95, pr = 0.7*. 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, *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. If the *pr* threshold is met, the second *cf* threshold is unnecessary, because it already met the first threshold. Accuracy was 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 may lower due to the small size of our subsets, or by forcing a consensus out of intentionally uncertain probability distributions, in which case accuracy is likely not the best measure of performance.

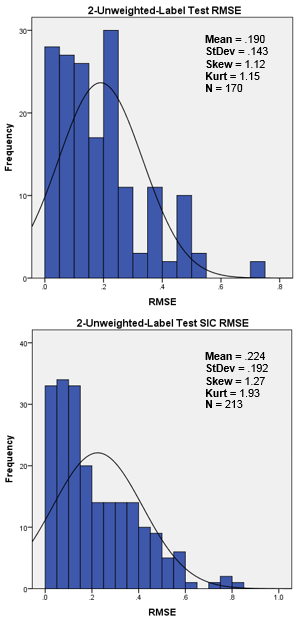
We calculated the RMSE between the actual and predicted distributions, which helps us assess the BDT’s ability to predict the correct distribution without forcing consensus. For the 5-label distributions, SIC afforded a slight increase in accuracy for 5-label distributions, as well as a slight decrease in mean and standard deviation RMSE and a slight increase in kurtosis and skew of the RMSE distribution, shown in Fig. 1.





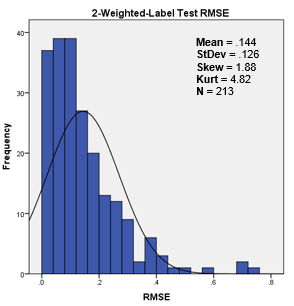
1. Distribution of RMSE between actual and predicted test labels for the original BDT model (top) and the BDT SIC model (bottom) for the 5-label distributions*.* The both BDT’s were run with np = 24, nc = 12, dmax = 12, k = 4, and the BDT SIC was run with cf = 0.95, pr = 0.5.

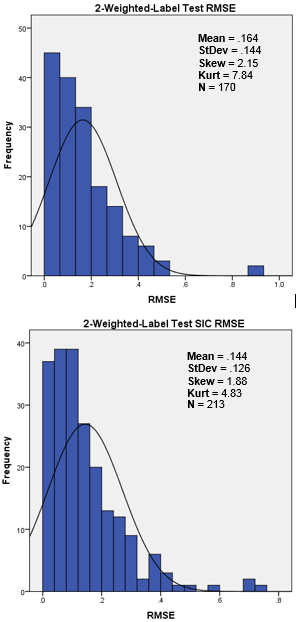




1. Distribution of RMSE between actual and predicted testing labels for: the original BDT model (top) and the BDT SIC model (bottom) with the 2-unweighted label distribution. Both BDT models were run with np = 28, nc = 14, dmax = 12, k = 4, and the BDT SIC model was run with cf = 0.95, pr = 0.7.

Results for unweighted 2-label models are shown in Fig. 2, and weighted 2-label models are shown in Fig. 3. 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.





1. Distribution of RMSE between actual and predicted testing labels for: the original BDT model (top) and the BDT SIC model (bottom) with the 2-weighted-label distribution. Both models were run with np = 28, nc = 14, dmax = 12, k = 4, and the BDT SIC model was run with cf = 0.95, pr = 0.7.

# Conclusions

Selective iterative classification may be a valuable tool to boost the accuracy of a probabilistic classifier, provided its classes are distinct enough.

Weighted reduction of the label set may also be worth further investigation, as it reduces the overall error in predictions with this type of distributed uncertainty.

Both of these approaches can be valuable assets in enhancing CAD system performance. By applying these techniques, CAD system accuracy with uncertain training data can be significantly improved, which will encourage adoption of the technology in clinical settings.

##### Acknowledgments

What should we add here???

##### References

1. K. Drukker, C. A. Sennett, and M. L. Giger, "Automated method for improving system performance of computer-aided diagnosis in breast ultrasound," *IEEE Transactions on Medical Imaging*, vol. 28, no. 1, pp. 122–128, Jan. 2009.
2. S. Jagdale, M. H. Kolekara, and U. P. Khot, "Smart sensing using Bayesian network for computer aided diagnostic systems," *Procedia Computer Science*, vol. 45, pp. 762–769, 2015.
3. M. K. Tafreshi *et al.*, "Su1662 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," *Gastrointestinal Endoscopy*, vol. 79, no. 5, pp. AB357–AB358, May 2014.
4. S. G. Armato *et al.*, "Lung image database consortium: Developing a resource for the medical imaging research community1," *Radiology*, vol. 232, no. 3, pp. 739–748, Sep. 2004.
5. R. Ochs, H. Kim, 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.," *Medical Imaging: Computer-Aided Diagnosis*, 2007.
6. Z. Elouedi, K. Mellouli, and P. Smets, "Belief decision trees: Theoretical foundations," *International Journal of Approximate Reasoning*, vol. 28, no. 2-3, pp. 91–124, Nov. 2001.

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

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

[9] J. Whitehill, T. Wu, J. Bergsma, J. Movellan, and P. Ruvolo, "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise," *Neural Information Processing Systems*, 2009.

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

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

[12] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, Oct. 2009.