An Evaluation of Consensus Techniques for Diagnostic Interpretation

Jake N. Sauter, Victoria M. LaBarre, Jacob D. Furst, Daniela S. Raicu
College of Computing and Digital Media, DePaul University, Chicago, IL, US

Abstract

Learning diagnostic labels from image content has been the standard in computer-aided diagnosis. Most computer-aided diagnosis systems use low-level image features extracted directly from image content to train and test machine learning classifiers for diagnostic label prediction. When the ground truth for the diagnostic labels is not available, reference truth is generated from the experts' diagnostic interpretations of the image/region of interest. More specifically, when the label is uncertain, e.g. when multiple experts label an image and their interpretations are different, techniques to handle the label variability are necessary.

In this paper, we compare three consensus techniques that are typically used to encode the variability in the experts' labeling of the medical data: mean, median and mode, and their effects on simple classifiers that can handle deterministic labels (decision trees) and probabilistic vector of labels (belief decision trees). Given that the NIH/NCI Lung Image Database Consortium (LIDC) data provides interpretations for lung nodules by up to four radiologists, we leverage the LIDC data to evaluate and compare these consensus approaches when creating computer-aided diagnosis systems for lung nodules.

First, low-level image features of nodules are extracted and paired with their radiologists' semantic ratings (1= "most likely benign,"… , 5 = "most likely malignant"); second, machine learning multi-class classifiers that handle deterministic labels (decision trees) and probabilistic vector of labels (belief decision trees) are built to predict the lung nodules' semantic ratings. We show that the mean-based consensus generates the most robust classifier overall when compared to the median- and mode-based consensus. Lastly, the results of this study show that, when building CAD systems with uncertain diagnostic interpretation, it is important to evaluate different strategies for encoding and predicting the diagnostic label.

1. Introduction

Lung cancer is the leading cause of cancer death and the second most common cancer among both men and women[1] **[1]**, causing a critical need to explore ways to diagnose malignant lung nodules early on. For that reason, under the National Cancer Institute's mandate, five academic institutions (Cornell University; the University of California in Los Angeles; the University of Chicago; the University of Iowa; and the University of Michigan) created the Lung Image Database Consortium (LIDC) to encourage the advancement of Computer-Aided Diagnosis (CAD), which assists radiologists in identifying cancer nodules within lung radiographs **[2]**. The developers of the LIDC warned future CAD developers to appreciate the substantial variability of ratings (such as whether an individual image had lung nodules, the number of lung nodules and whether the nodules were benign or malignant) among the panel of four radiologists that had analyzed the images for the database and established a "truth" baseline.

Computer-aided diagnosis (CAD) systems have been developed to assist radiologists by providing an additional opinion. To create these CAD systems, we require label data from radiologists representing their diagnosis interpretations. However, annotations from multiple radiologists are often uncertain because of the subjectivity of the interpretation process and the variability among annotators. We propose a study of different approaches to reaching a consensus label that minimizes false positive and false negative rates and maximizes true negative and true positive rates. More specifically, we propose that the mean consensus of a list of votes for the classification of an object is the best consensus to form a ground truth with. The premise of this hypothesis is that past methods of evaluating Computer Aided Diagnosis tend to make use of performance metrics such as H-Measure, Area Under the Receiver Operator Characteristic Curve, and similar metrics, and these metrics work well with binary class outputs, however it is intuitive to recognize

that data may be lost by compressing the probabilistic distribution into only two classes rather than a distribution.

Comparing distributions not only allows data to be retained, but also allows for more intuitive evaluation of outputs, which is especially important when lacking the presence of a ground truth. Moving our focus away from adjusting metrics, consensus techniques are innately capable of comparing distributions and are not restricted to binary class outputs. To clarify, using a consensus to best represent the classification is more informative than presenting it as a binary classification, as the most likely class (1-5) will be presented, without the struggle of matching a probability distribution completely. This gives relevance to testing how consensus techniques, rather than metrics, affect the behavior of classifiers, specifically Belief Decision Trees. Understanding the effects of consensus techniques (mean, median and mode) has the potential to assist future studies when developing or testing metrics through comparison.

2. Related Works

In the past, most metrics used to quantify the variance within the LIDC utilized binary, class outputs, compressing the probabilistic distribution into only two classes rather than a distribution has the potential to cause a loss in significant data. For this reason, some researchers took the Area Under the Receiver Operator Characteristic Curve and developed a new metric called the Area Under the Distance Threshold Curve, which allows for comparison of distributions.

One study, **[3]** specifically quantified the variance of the radiologist's ratings based on the differences in how the radiologists outlined where they believed a cancer nodule was within a radiograph by using a P-mapping for each pixel. This study confirmed the warnings mentioned by the creators of the LIDC database concerning high variance in radiologist's contours of LIDC data, especially in certain images.

Due to this variance issue, several research studies have been conducted in attempts to quantify the credibility of the data (the radiologist's ratings) and the confidence of semantic features of lung nodules **[3]**. An important semantic feature captured by this database was the malignancy rating on a five-point scale (1 being most benign and 5 being most malignant) by a radiologist. The concept behind this line of research is that a radiologist is unlikely to seek CAD help if he or she is confident that a lung cancer nodule is benign or malignant because then he or she knows how to handle that case. However, CAD can be used to identify more difficult to diagnosis cases, such as the cases where the panel of four radiologist disagreed the most. If a CAD algorithm can identify a difficult case where the radiologists are unsure of a lung cancer rating for a nodule, then hypothetically, this is the image that a radiologist should devote more of their limited time to.

In the past, researchers have off-shot from more traditional evaluation methods to newer methods, such as the Area Under the Distance Threshold Curve (AUCdt) which evaluates probabilistic multi-class classifications with the intention of quantifying uncertainty **[4]**. In that same study, researchers concluded that an ensemble of classifiers significantly improved performance over just a single classifier. The previously mentioned study was expanded in another study **[5]** which utilized the Area under the Distance Threshold Curve to evaluate probabilistic multi-class systems and delivered a five-label distribution. This supports our study's reasoning behind using a distribution output instead of a binary output classifier. Our

study proposed the use consensus of the probabilistic distribution instead of distribution similarity metrics to later compare to previous and future studies.

Another relevant study specifically used Belief Decision trees with a distribution and compared traditional Belief Decision trees using the Area under the Distance Threshold Curve **[6]**. Researchers conducting this study noted that AUCdt showed to be an improved measure and that most errors were caused by low confidence. In conclusion, this study found that their predictions agreed with a Benign diagnosis more often than radiologists, which means that although the number of false positives was reduced, the number of false negatives increased **[6]**. This is expected, as the inherent drawback of decreasing false negatives is to gain a higher likelihood of more false positives and vice versa. As further support for the use of belief decision trees, one study used a multiple-label classification algorithm based on belief decision trees where each nodule was viewed as four distinct instances when training the classifier **[7].** This same study also evaluated their results using an AUCdt measure and used multiple-label classification algorithms to display diagnosis when no ground truth was available. A similar study predicted distributions using multi-label classification implemented with belief decision trees and concluded that multiple-label classification algorithms are an appropriate method of representing the diagnoses of multiple radiologists on lung CT scans when ground truth is unavailable **[8]**.

Focusing away from belief decision trees and onto the topic of malignancy ratings, one study focused on improving how a classifier determines whether hard-to-distinguish cases were benign or malignant. This study obtained their data from a previous data set similar to the LIDC, and if any of the radiologists marked the breast mass as hard to distinguish, also referred to as a malignancy rating of three, then it was included in the data set. The more relevant parts of this study to mention are that for each lesion, blinded retrospective interpretation was performed by five radiologists to provide BI-RADS category and that there were 69 masses (21 malignant and 48 benign masses) that were assigned to BI-RADS category 3 by at least one of five radiologists included in this study **[9].**

In another study that focused on classifying malignancy ratings, researchers used a weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics. The same study used the term “vote vector” for a vector with semantic meaning extrapolated from the LIDC data, which was also used to train a classifier to predict malignancy **[10]**. Other studies also used the LIDC to train a classifier to detect malignancy. One study noted that associated truth files could be referenced, and all other cases disregarded, however they only exist for some small percent of the cases, rendering most of the dataset useless. This is a common problem when quantifying variance within the LIDC **[11]**.

In the same line of thought, another study aimed to adjust the amount of variance within the dataset by addressing the radiologists directly with the intention of improving the radiologists’ ratings, rather than create a consensus using a classifier. In this study, researchers set out to see if the quality of the original four radiologists’ readings improve after participating in a program meant to help train them. This study concluded that the Radiologists who undertake the BREAST program demonstrated significant improvements in test-set performance during a 3-year period, the significance of which highlights the value of ongoing education using of test-sets **[12]**.

Although addressing variance between radiologists directly by sending radiologist through a specific program clearly has the potential of lowering the amount of variance, it can be inferred that variance will still occur since not all radiologists agree. This gives relevance to drawing a consensus from the data and using a classifier. In the past, some studies used LIDC to train a classifier to detect malignancy, such as is denoted in the study **[13]**. In the previously mentioned study, researchers were able to get a consensus from the separate radiologist ratings using image acquisition. The most relevant aspect of this study with respect to the study outlined in this paper that that the previously mentioned study used the mode of the radiologist ratings, and if mode was not present, the floor of the median was taken. The use of mode implies that this previously mentioned study chooses a single consensus to access a ground truth from a LIDC sourced dataset. For this reason, our study, outlined in this paper, is relevant for comparison because it uses three consensus techniques: mean, median and mode, rather than mode alone.

After pointing out the basis and relevance of our study, it is important to note that our study, outlined in this paper, is directly built from the ground work laid out by a past study **[14]** from the same lab at DePaul University. This previous study used Belief Decision Trees with probabilistic outputs and used Area Under the Distance Threshold curve as an accuracy measure **[6].** Another study, also from DePaul University's lab, was concerned with quantifying disagreement and the lack of ground truth in the LIDC. This in-house study noted that no type of analysis has been done for the type of disagreement that the LIDC is dealing with, being difference in semantic ratings of medical diagnosis severity, which gives relevance to the study outlined in this paper. **[15]**

All the previously mentioned studies support our experimental choice that picking a best consensus technique to represent the classification is more informative than presenting the output as a binary classification, as the most likely class (1-5) will be presented, without the struggle of matching a probability distribution completely.

3. Methods

Our proposed methodology to handle discord among radiologists in the semantic ratings (only malignancy will be present in this study) in the LIDC comes in 3 phases. First, image features of nodules are extracted and correlated to their semantic ratings of malignancy. Second, a belief decision tree is built to predict the semantic rating of the labelled training data. Third after training was completed, the training and testing set is classified by the trained BDT and the basic belief assignments of the cases classified in the testing are evaluated and compared based on true negative/positive rates and false negative/positive rates.

3.A. The Lung Image Database Consortium (LIDC) Data

The LIDC dataset **[2]** (available at http://ncia.nci.nih.gov) is a diverse collection of Computed Tomography (CT) scans of 1,000 patients interpreted by up to four radiologists. Each radiologist outlined a boundary for the nodule or nodules present in the scan, and provided subjective ratings on various semantic characteristics (such as texture, margin, spiculation, lobulation and malignancy) for each nodule as a whole. The CT scans available in the LIDC come differing points in time and imaging machines, leading to varying slice thicknesses. In total 2600 nodules were outlined and annotated, however regarding the malignancy of a nodule agreement only occurred in 25% of the cases. This observation of uncertainty of ground truth yields the necessity for the analysis of the present label data. The challenges of the LIDC data include the disagreement among radiologist annotators, the multi-class label (for example, there are five ratings for malignancy rather than the traditional two class problem - malignant versus benign), and lack of ground truth (whether the annotated nodule was later determined to be malignant or benign). These challenges

make the LIDC data a good candidate to study behavioral changes of classifiers when the reference truth is calculated in various ways.

While in this study we focus only on the malignancy ratings, the same approach can be applied to the other semantic characteristics (e.g. degree of spiculation, lobulation, and texture). Further, to analyze the variability among annotators, we considered a subset of the LIDC consisting of only the 809 nodules out of the 2,600 nodules for which all four radiologists outlined and annotated nodules in the image and provided a rating for malignancy. For further balancing of this dataset, cases that the consensus of the radiologists was uncertain (3) were under-sampled, removing around 150 cases. Also, to increase the amount of certain cases, around 50 random cases were duplicated, leading to a more balanced data set with 850 cases, similarly to what was done in **[14]**, aiding the training of most classifiers. A necessary distinction is that only nodules with a minimum size of 5x5 pixels, or 3x3 mm were analyzed, allowing meaningful texture data to be extracted from nodules. For each nodule 63 two-dimensional image features were extracted from the largest radiologist outline associated with the nodule. After extraction, a vector representation is taken of each nodule with 63 image features and 7 annotations from each radiologist. More information on the features are ratings can be found in **[15]**.

3.B. Belief Decision Trees

Belief Decision Trees (BDTs) are constructed much like traditional decision trees except with the caveat of the decision to split at a node is based on averaging the pignistic probabilities, being Basic Belief Assignments (BBAs) in BDTs **[17]**. Our study is in different consensus techniques, which produce a consensus rating from the four radiologists' ratings, which is then converted into a BBA, leading to the production of different BBAs from different consensus techniques. Calculating BBAs is usually a more complicated process, however the inherent properties of the LIDC allows for a simpler process in this study. We can eliminate the possibility of having one or more "true" labels because every radiologist can only choose one malignancy rating for each case. Also, pure uncertainty is not represented in the LIDC: a rating of 3 represents an equal probability of malignant or benign. These traits of the LIDC allow for a simplification of the calculation of ratings to a probability distribution **[8]**. It is important to note that a 70/30 test train split was used.

To decide if and how to split at a node, the average pignistic probabilities of the parent and child nodes can be used to calculate the information gain, which is then used to decide the optimal split with respect to each possible feature and threshold value in the dataset. A gain ratio that rewards equally distributed child subset splits was taken advantage of to produce a more balanced tree, similarly done in **[14]**. The low-level image feature that achieves the maximum gain ratio for the corresponding split is then selected at the current node to perform the partitioning of the cases. Four separate stopping criteria were applied, being that if any or all cases were met, the current node would not be split and would be a leaf node. The stopping criteria are as follows, the gain ratio of a split at a node is 0, there is no split that can be made which will result in acceptable numbers of cases at the parent and child nodes, all the BBA's at the node are equivalent, or all features have already been used to split **[17]**.

3.C. Low-level Image Feature and Consensus-based Label Extraction

Based on our previous work **[8]**, we extracted 64 low-level image features for each nodule instance delineated by the radiologists' largest outline across all slices in which the nodule appeared **[18]**. These image features encode the lung nodule intensity, size, shape, and texture.

The malignancy label has five ratings (1= "most likely benign," … , 3 = "indeterminate," … , 5 = "most likely malignant) which makes the classification a multi-class problem. Furthermore, given the variability

in the radiologists' interpretation, for the same nodule, there could be multiple ratings. Therefore, we encode the malignancy label using a vote vector which is further converted into a probabilistic label vector (PLV) which represents the probability distribution over the five possible class ratings of malignancy for that specific nodule.

Three methods of calculating a zero-initialized PLV, denoted as *P,* from a vote vector *V* were tested, what we call mean, median and mode consensus techniques. These techniques generate a consensus label *C*, which is then converted into *P.* The mean vote vector consensus technique is applied by first arriving at *C* by the standard definition of mean. This *C* is converted to a probability distribution over the five possible classes. In a similar fashion, the median- and mode-based consensus label *C* are created for the same nodule diagnostic interpretation. For example, if the radiologists' rating for a certain nodule was *V*= [2, 2, 3, 4], the mean of this vote vector is *C*=2.75, leading to a *P* vector defined as [0, .25, .75, 0, 0] with a higher probability assigned to 3 than 2 given that the mean is closer to 3 than to 2. Similarly, for median and mode conversion of *V*, *C* = 2.5 and 2, respectively, which will further produce P = [0, .5, .5, 0, 0] for the median label and P= [0, 1, 0, 0, 0] for the mode label.

3.C.1 Actual Label Conversions

The semantic ratings of malignancy aforementioned were collected for each nodule, from four radiologists. Each of the radiologists rated the malignancy of a nodule on a one to five scale, with a rating of one being completely benign and a rating of five being completely malignant. These ratings were concatenated into a list form ("**vote vector**") for each nodule, before one of three possible consensus techniques was applied to get a consensus label for that nodule. This vote vector is later converted into probabilistic label vector ("**PLV**") which represents a probability distribution over the five possible class ratings of malignancy level for that specific nodule.

3.C.2 Predicted Label Conversions

Three consensus methods for calculating a semantic class rating *S* from a BBA *B* were applied. These three techniques were synonymous calculations to the consensus techniques applied to produce PLVs. The mean BBA consensus technique sums the product of each class with its corresponding probabilistic weight.

The median BBA consensus technique works in the following way, a running sum *R* is taken of *B*, keeping track of the percentage of the probability distribution contained within and below the current index. The first index in which the sum at is greater that 0.5 is selected, indicating that the current index is the index where the 50th percentile of the distribution occurs. The mode BBA consensus technique simply selects the most probable class in *B*.

3.C.3 Consensus

We name our approach of encoding the uncertainty of the label *consensus PLV* (Figure 1) because it first creates the consensus and then it encodes it as a probabilistic vector of labels. This is different from previous approaches **[4]**, including ours, in which we used a *direct PLV approach* in which the vector V is converted to P directly without considering consensus. For example, the direct PLV approach will convert V into P based on the probabilities of each rating (*P*= [0, .5, .25, .25, 0]).

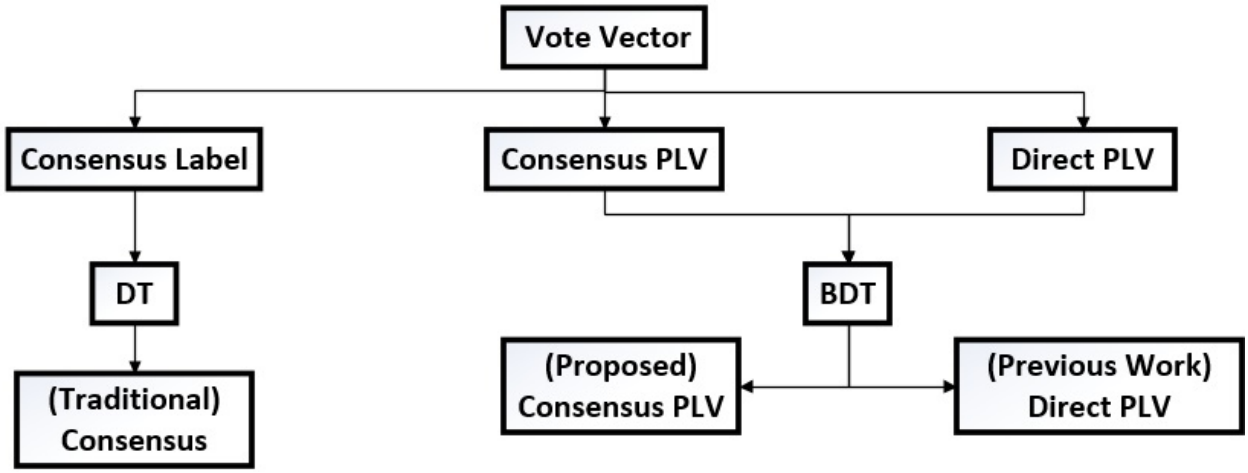

Figure 1: Diagram of the different methods to evaluate consensus-based classifiers

3.D. Classification Models

We analyze the impact of encoding the reference truth through consensus-based approaches using Belief Decision Trees (BDTs). BDTs inherently model uncertainty through predictions being probabilistic distributions over the possible classes. This, however, requires extra processing because the vote vector *V* is converted to consensus label *C*, then into a probabilistic label vector *P* for a target label to train the BDTs, and then the predicted *P* is converted back to predicted *C* for classification performance assessment. Therefore, we compare the results for the belief decision trees with those obtained using simple decision trees **[17]** trained and tested on the consensus-based label C (Figure 1).

The construction of the BDTs **[17]** is similar to the traditional decision trees with the exception that the decision to split at 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. 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, four stopping criteria were used: 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, all the BBA's at the node are equivalent, or all features have already been used to split **[16]**.

4.Results

4. Accuracy

The actual and predicted consensus (mean, median and mode) of the predicted and actual PLVs of each case were compared using accuracy.

	Predicted Mean	Predicted Median	Predicted Mode
Actual Mean	**46. 099**	**46. 099**	**46. 099**
Actual Median	40.425	40.425	40.425
Actual Mode	30.858	30.858	30.858

Table 1: Accuracy Results

From the accuracy results we were able to make a key observation, being that the mean consensus technique allowed the classifier to reach the highest level of accuracy. However, this was not just for the mean input and output consensus pair, but for all output consensus techniques paired with the mean input consensus technique.

4.B True Positive/Negatives and False Positives/ Negatives

In Figure 2 depicted below, true positive (TP), correctly classified indeterminate (TI), true negative (TN), false positive (FP), incorrectly classified indeterminate (FI), false negative (FN) rates are compared across consensus-based techniques and classification models. The grouping of bars on the left represents the actual negative cases and the grouping on the right represents the actual positive cases. Within the groupings, spans at the top of the bars indicating results from the DTs can be found on the left and results from the BDTs can be found on the right can be found. The red regions of the bars represent incorrectly predicted category, grey regions represent indeterminate predicted category (predicted as a rating of 3), and green regions represent correctly predicted category. Included at the right of these two groupings are rates calculated from directly comparing the distribution of the BDT without forming a consensus label, to compare how forming a consensus affects these rates.

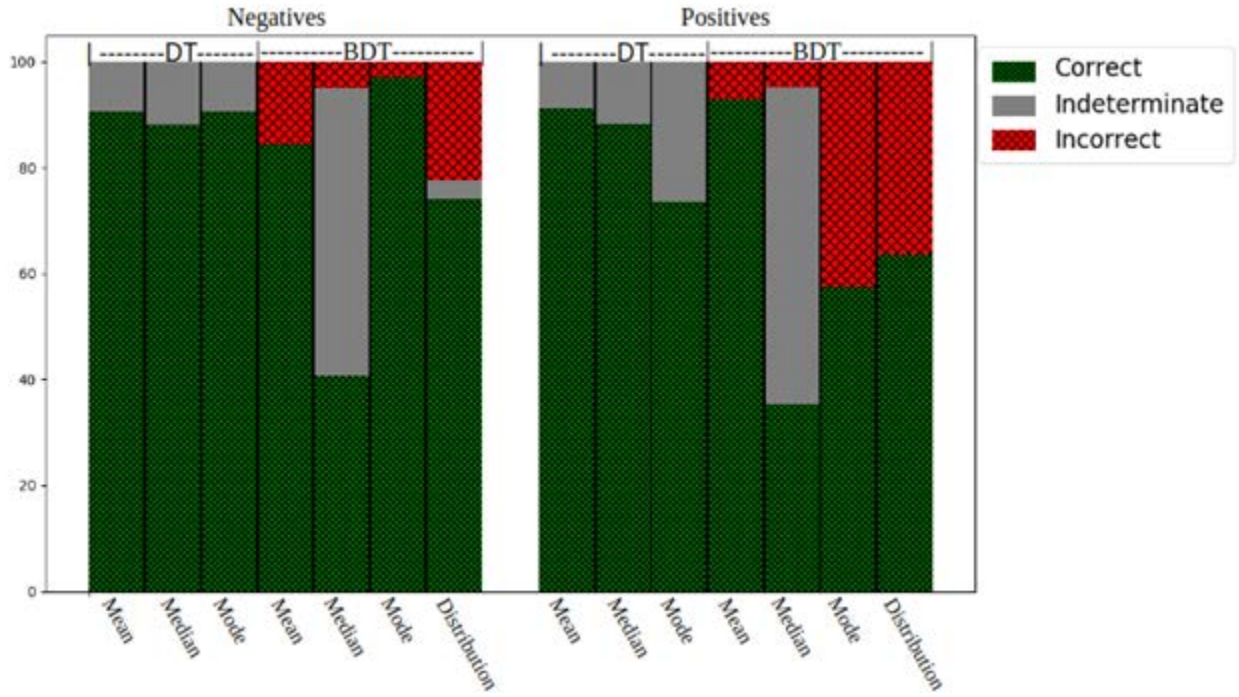

Figure 2: True Positive/Negative and False Positive/Negative Rates, includes Indeterminate Rate

It is important to specify that labels predicted as a three (neither malignant nor benign) are calculated separately as the "Indeterminate Rate." This means that the total true positive/negative or false positive/negative rates do not sum to a total of 100 percent.

From Figure 2, we observed behavioral traits for each consensus technique that led us to conclude that mean is the best consensus technique. (It is important to note here that this study was mainly interested on how the behavior of the classifier changed when different consensus techniques were chosen). More specifically, mean was the best consensus technique for avoiding both false positives and false negatives,

and with improved accuracy of the classifier, predicted ratings will vary more, becoming closer to desired classifications. In Figure 2, notice that mean used within a BDT had less misclassifications than the distribution, even though the mean had a higher number of misclassifications of negatives than median and mode within a BDT. Within the DT, mean had higher or equal correct classifications of both negatives and positives when compared to median and mode, also within the DT. Therefore, due to the similar behavior of mean compared across the BDT and the DT, mean was concluded to possess the most optimal behavior as a consensus technique.

With regards to the correct classifications of positives, mean was greater than median, which had the highest number of indeterminates and higher than mode which had the highest number of false positives within the BDT. Within the DT, mode continued to have the least amount of correctly classified positives, and although median had less indeterminates, mean still had the highest number of correct positive classifications. It can be inferred from the previous statement that mode tends to have a higher false positive rate than other consensus techniques when using a BDT or DT. On the other hand, mode was found to have the highest correct classification of negatives within the BDT, compared to mean which had the highest number of misclassifications and median which had less misclassifications than mean but a drastically higher number of indeterminates.

Overall, median appeared to possess undesirable behavior in both the BDT and DT as it consistently had the highest number of misclassifications, except for positives within the DT where its number of misclassifications fell behind mean but was still higher than mode. Lastly, all three of the consensus techniques (mean, median and mode) had more desirable behavior than taking a direct distribution across both a BDT and DT. Therefore, using a consensus technique in general is preferable.

Conclusions

In most studies that explore consensus, choosing a consensus for a list of votes is necessary but the method of consensus has not been deeply explored **[13].** This study is the first to perform a comparison of classifiers' behavioral changes with use of different consensus techniques to form a reference truth. This study also investigates the comparison of a classifier with a single class output against a classifier that outputs a probabilistic distribution over possible classes.

After comparing results from a classifier with and without consensus integrated in the label, we found that a consensus produced more robust results. Mean-based consensus has the highest ratio of positive responses to negative responses among the three consensus techniques over both classifiers. We also found that the model that was trained on a consensus PLV performed better than the model trained on a direct PLV, because the prediction indicating the correct single label was our qualifier for accuracy. We should note that comparing distributions not only allows data to be retained, but also allows for more intuitive evaluation of outputs, though a tradeoff of robustness to information loss is present. This tradeoff is especially important to consider when lacking the ground truth. Future research into BDTs and their probabilistic vector outputs will be pursued, possibly utilizing AUCdt **[5]** to aid in the evaluation of a distribution-based classifier rather than consensus-based classifier.

Acknowledgements

This work was supported by the National Science Foundation under Grant No.1659836. We would also like to thank DePaul University's College of Computing and Digital Media for providing us with the LIDC image features for each nodule.

References

[1] American Cancer Society. "Key Statistics for Lung Cancer." American Cancer Society. American Cancer Society, 5 Jan. 2017. Web. 17 July 2017.

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

[3] Zinoveva O., Zinovev D., Siena S. A., Raicu D., Furst J., and Armato III S., "A texture-based probabilistic approach for lung nodule segmentation", Lecture Notes in Computer Science, International Conference on Image Analysis and Recognition, 2011

[4] Zinovev, Dmitriy, Jacob Furst, and Daniela Raicu. "Building an Ensemble of Probabilistic Classifiers for Lung Nodule Interpretation." *IEEE* (2011): 155-61. Web. 17 July 2017.

[5] Williams, Sydney, et al. "Area under the Distance Threshold Curve as an Evaluation Measure for Probabilistic Classifiers." SpringerLink, Springer, Berlin, Heidelberg, 19 July 2013, link.springer.com/chapter/10.1007/978-3-642-39712-7_49.

[6] Zinovev, Dmitriy, Yujie Duo, Daniela Raicu, Jacob Furst, and Samuel Armato. "Consensus Versus Disagreement in Imaging Research: a Case Study Using the LIDC Database." *J Digit Imaging*(2011): n. pag. Web. 17 July 2017.

[7] Zinovev, Dmitriy; Feigenbaum, Jonathan; Raicu, Daniela; and Furst, Jacob, "Predicting Panel Ratings for Semantic Characteristics of Lung Nodules" (2010). Technical Reports. Paper 18. http://via.library.depaul.edu/tr/18

[8] Zinovev, D., 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): n. pag. Web.

[9] Moon, Woo Kyung, Chung-Ming Lo, Jung Min Chang, Chiun-Sheng Huang, Jeon-Hor Chen, and Ruey-Feng Chang. "Quantitative Ultrasound Analysis for Classification of BI-RADS Category 3 Breast Masses." *Journal of Digital Imaging* 26.6 (2013): 1091-098. Web.

[10] Kaya, Aydın, and Ahmet Burak Can. "A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics." Journal of Biomedical Informatics 56 (2015): 69-79. Web.

[11] Sun, Wenqing, Bin Zheng, and Wei Qian. "Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis." *Computers in Biology and Medicine* (2017): n. pag. Web.

[12] Suleiman, Wasfi I., Mohammad A. Rawashdeh, Sarah J. Lewis, Mark F. Mcentee, Warwick Lee, Kriscia Tapia, and Patrick C. Brennan. "Impact of Breast Reader Assessment Strategy on mammographic radiologists test reading performance." *Journal of Medical Imaging and Radiation Oncology* 60.3 (2016): 352-58. Web.

[13] Filho, Antonio Oseas De Carvalho, Aristófanes Corrêa Silva, Anselmo Cardoso De Paiva, Rodolfo Acatauassú Nunes, and Marcelo Gattass. "Computer-Aided Diagnosis of Lung Nodules in Computed Tomography by Using Phylogenetic Diversity, Genetic Algorithm, and SVM." *Journal of Digital Imaging* (2017): n. pag. Web.

[14] Affenit, Rachael N., et al. "Building Confidence and Credibility into CAD with Belief Decision Trees." Medical Imaging 2017: Computer-Aided Diagnosis | MI17 | SPIE Proceedings | SPIE, International Society for Optics and Photonics, 3 Mar. 2017 spiedigitallibrary.org/proceeding.aspx?articleid=2609070.

[15] Zinovev, Dmitry et al. "Predicting Radiological Panel Opinions Using a Panel of Machine Learning Classifiers." *Algorithms* 2 (2009): 1473-1502.

[16] Horsthemke, William H., Daniela S. Raicu, and Jacob D. Furst. "Evaluation Challenges for Bridging Semantic Gap." *International Journal of Healthcare Information Systems and Informatics* 4.1 (2009): 17-33. Web.

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

[18] Riely, Amelia, et al. "Reducing Annotation Cost and Uncertainty in Computer-Aided Diagnosis through Selective Iterative Classification." Medical Imaging 2015: Computer-Aided Diagnosis | MI15 | SPIE Proceedings | SPIE, International Society for Optics and Photonics, 20 Mar. 2015, proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=2211245.