Determining Reliability of Computer Aided Diagnosis Outcomes Involving Uncertainty and Multiple Rater Disagreement

Yenong Du[1]

1.College of Computing and Digital Media, DePaul University, Chicago, Illinois 60604, USA

emmayduyu@gmail.com

ABSTRACT. Using the Lung Image Database Consortium (LIDC) computed tomography (CT) scans and data of lung nodules, we implement an adaptation of a k-Nearest Neighbors (KNN) algorithm to predict probabilistic vectors for a multi-class, multi-label learning (MLL) problem to quantify the disagreement between a variable number of multiple raters and data uncertainty due to the use of a subjective ("semantic") rating system. We show progressive improvement in the predictive power of each model as we first capture agreement, certainty, then apply a cascading classifier, and finally address target class imbalance using Synthetic Minority Over-sampling Technique (SMOTE). We implemented two evaluation metrics - Mean Standard Error (MSE) and Area Under the Distance Threshold Curve (AUCdt) using Jeffrey's Divergence distance. The experimental outcome shows the validity of embracing both disagreement and uncertainty within multiple rater data as evidenced by the gains in evaluation metric values.

Keyword: multipliable classification, probabilistic classifier, AUCdt, K-Nearest neighbor

1.INTRODUCTION

Lung cancer is the deadliest form of cancer in the United States accounting for more deaths than colon, breast, and prostate cancers combined (American Cancer Society 2018). Early detection is one of the most effective ways to improve chances of survival. As part of the solution to improve detection and diagnosis, the Lung Image Database Consortium image collection (LIDC-IDRI) was developed to facilitate a publicly accessible clinical thoracic computed tomography (CT) scan repository to empower the medical imaging research community to develop computer aided diagnostic (CAD) methods for lung nodule detection, classification, and quantitative assessment.

The Lung Image Database Consortium (LIDC) was developed to facilitate a publicly accessible "well-characterized repository of CT scans" for the medical imaging research community to allow further development of "CAD methods for lung nodule detection, classification, and quantitative assessment" (Armato et al. 2011). It contains CT scans of lung nodules annotated by up to four radiologists and has enabled the academic community to develop and improve methodologies and algorithms toward the detection and diagnosis of cancerous nodules within the lung. We leveraged the data from the LIDC to focus specifically on CAD – a system designed to provide radiologists with an automated "second reader" to increase accuracy (and speed of treatment) of potentially harmful anatomical structures. Research efforts to accomplish this have been earnestly pursued since the 1980s (Kunio Doi, 2007), yet further opportunities for improvement exist.

Inspection of the LIDC dataset reveals that each CT scan slice or image is accompanied by a series of image measurements and set of semantic ratings provided by up to four radiologists from a pool of twelve participating radiologists. These subjective ratings describe qualitative semantic features observed in lung nodules throughout various slices within a CT scan. A final rating captures the radiologists' determination of likelihood of malignancy. Because nodule characteristics are complex, and diagnosis is not always clear cut, the ratings between experts do not always agree. The presence of rater disagreement and uncertainty of a subjective rating system is readily found in diagnostic imaging research literature (Khaw et al. 2016, Perez-Gomez et al. 2012, Yamane et al., Spayne et al. 2012) and imposes critical influence on the complexity of this research area (Zinovev et al. 2012).

Because of this complexity, research efforts involving this dataset have heavily relied on techniques that reduce the ambiguity caused by disagreement and uncertainty of rater data. One common approach for reducing multiple ratings is use of a consensus method, such as using the mode or mean rating. However, in doing so, nuances of information captured by individual rater scores can be lost through this forced aggregation even though several researchers have found that consensus rarely occurs even among expert radiologists (Zinovev et al. 2012, Jarvik et al. 2009). Other research has also taken an approach to qualify

cases with a minimum number of raters – some requiring at least two raters to agree and others selecting only cases with four raters (Opfer et al. 2007, Reeves et al. 2007, Yung et al. 2017).

Limitations and the inherent loss of information of these approaches are widely recognized. With few exceptions (Kaya et al. 2015), pioneering efforts to leverage the full amount of information present in multi- label datasets take form through problem transformation approaches. Of these approaches, binary relevance and label powerset techniques are gaining attention (Tsoumakas et al. 2007). These techniques attempt to minimize information loss, but also involve challenges. For example, binary relevance assumes label independence that often – and certainly in the case of the LIDC data – does not exist in real world data. The loss of this correlation between labels is a significant limitation of these methods. Label powersets mitigate this disadvantage by preserving inter-label relationships, but because of the potential for exponential increases in the number of potential label sets, the scalability – and therefore computability – of this approach is an inherent weakness.

To capture the nuanced information present in the semantic ratings, we needed to address two challenges. One challenge is that of *disagreement* and the other of *uncertainty*. Disagreement occurs when at least two raters rate a semantic feature differently, even if they might perceive that rating to describe the same observation. Further, there is no guarantee that either rater is "correct" as the ratings are subjective. This idea illustrates the meaning of uncertainty. Additionally, the data also contains a variable number of participating ratings. We operate on an assumption that more participating raters are more likely to reach a "ground truth," similarly to voting by majority.

In order to address these challenges, we quantified these concepts as new features through calculations involving the number of raters participating, number of raters in agreement, and the total number of potential raters. We then built models to classify probabilistic vectors to solve this multi-label, multi-class problem. Thus, our research question was to determine whether we could improve classification performance of CAD outcomes by embracing both uncertainty and disagreement in a probabilistic vector classification model.

The remainder of this paper is organized as follows: Section 2 describes related works; Section 3 discusses materials and methods including the LIDC dataset, data transformations undertaken, methodologies deployed, and experimental design to address the research question; Section 4 and 5 provide results and the discussion of those results; Section 6 covers conclusions and future work; followed finally by references and appendices.

2.RELATED WORKS

Owing to its popularity among the medical diagnosis research community, the LIDC dataset has been used to develop several CAD systems. In these works, a combination of image features, semantic features, or both were used to develop methods for classifying malignancy diagnoses. While some works have embraced disagreement among the panel of raters, most works were based on a consensus approach, where feature aggregation is applied to reduce the research task to a lower dimensional problem. Yung et al. 2017 applied a consensus approach using the mode or ceiling of the mean of the semantic ratings, low image features, and synthetically generated instances via oversampling (SMOTE) to develop a random forest classifier.

Zinovev et al. 2009 developed a semi-supervised, probabilistic model using an ensemble of classifiers that focused on the agreement – or inter-observer variability – among the radiological panel of raters. Dhara et al. 2017 used a combination of 2D shape-based, 3D shape-based, 3D margin-based, 2D texture-based, and 3D texture-based features to develop a Support Vector Machine (SVM) to classify benign and malignant nodules. Other research has also focused on detection as well as diagnosis. Firmino et al. 2016 introduced a new system that combined both CAD detection (CADe) and CAD diagnosis (CADx) using 3D

segmentation images, 3D internal structure of lungs and detection candidates to develop a SVM detection and diagnosis classifier. This was significant because "combining CADe and CADx systems would improve the level of automation and efficiency for both detection and diagnosis with minimal user intervention."

Some works have embraced potential disagreement among radiologists. Sarfaraz et al. 2017 used a 3D Convolutional Neural Network (CNN) with transfer learning and risk stratification by embracing potential disagreement among raters to develop a 3D CNN. Zinovev et al. 2011a embraced disagreement by introducing the radiologists ratings into a probability vector of ratings while developing a belief decision tree classifier. Valizadegan et al. 2013 investigated different sources of disagreements by combining different ratings or labels from domain experts to obtain a consensus classification, as well as develop individual expert models. Their proposed model, the Multiple Experts Support Vector Machine (ME-SVM) provided better results by introducing diversity in the consensus model.

Additional research has focused on finding reliable measurement for multiple rater reliability for nominal data, i.e which coefficients and confidence intervals are appropriate. Zapf et al. 2016 compared Fleiss' kappa and Krippendorff's alpha statistical properties for assessing multiple rater reliability in varied situations. Zinovev et al. 2011b introduced Area Under the Distance Threshold Curve (AUCdt) as a MLL evaluation metric. Williams et al. 2013 further evaluated AUCdt by incorporating various distance measures such as City Block Difference, the Jeffrey Divergence, and Earth Movers Distance. The study "showed that the new quantities improved evaluation of multiclass probabilistic classifiers like ROC curve (AUC)". Kaya et al. 2015 used "votes and rules obtained from radiologist evaluations in a weighted rule-based method to predict malignancy by considering correlations between malignancy and other nodule characteristics and the agreement ratio of radiologist," although traditional classification evaluation metrics were used.

3.MATERIALS AND METHODS

In this work, we adapt a KNN classifier to accommodate probabilistic vectors to classify malignancy of CT Scan studies in a multiple rater CAD system. Our approach uses combinations of image features and semantic features that are calculated and weighted as probabilistic vectors for a base model and a cascading model classifier to tackle uncertainty and disagreement. SMOTE is also applied to the dataset to boost the imbalanced classes, which in theory will help to make the model more generalizable.

3.1. LIDC Dataset

We sourced our data from a previous study (Yung et al. 2017) accessible at https://www.dropbox.com/sh/zxq69reu1ez00xm/AACnbSXLHUg-Dqr6YYVwQl2Ba?dl=0. This dataset was derived from the LIDC dataset (Armata III et al. 2015) but has been preprocessed to contain only one slice per nodule containing the maximum intersection of pixels (area > 25 pixels) annotated by the expert raters. The final dataset contains 2,588 nodule slices with 64 image features (sec 3.2.1), nine semantic features (sec 3.2.2), and our additional calculated probabilistic vectors (sec 3.2.3). The following sections describe each set of features and our method for preprocessing data to capture disagreement and uncertainty. Including the transformed vectors, our final dataset contained 2,588 instances of 92 features containing 4,484 unique labels in the weighted malignancy vector resulting in a label cardinality of 1.73.

3.1.1. Image Features

The 64 extracted LIDC image features include the following continuous data: size features (7), shape features (8), intensity features (9), and texture features (40) (Table 1).

Table 1. LIDC Image Features

Size	Shape	Intensity
Area Convex Area Perimeter Convex Perimeter Equivalent Diameter Major Axis Length Minor Axis Length	Circularity Roughness Elongation Compactness Eccentricity Solidity Extent Radial Distance SD	Minimum, Maximum, Mean Intensity Intensity SD Minimum Intensity Background Maximum Intensity Background Mean Intensity Background SD Intensity Background Intensity Difference

Texture	**Harlick Features:** Calculated from co-occurrence matrices (Contrast, Correlation, Entropy, Energy, Homogeneity, 3rd Order Moment, Inverse Variance, Sum Average, Variance, Cluster Tendency, Maximum Probability)
	Gabor Features: Mean and standard deviation of twelve Gabor images (orientation = 0°, 45°, 90°, 135° and frequency = 0.3, 0.4, 0.5)
	Markov Random Fields (MRF) Features: Means of four response images (orientation = 0°, 45°, 90°, 135°) and variance response image

3.1.2. Semantic Features

Eight semantic features plus the semantic malignancy feature were quantitatively scored by a pool of twelve radiologists based on the evaluation of the CT scans (Table 2). Each nodule was rated by up to four radiologists from a pool of twelve.

Table 2. LIDC Semantic Features

Semantic Feature	Description	Semantic Feature	Description
Subtlety Difficulty of detection	1.Extremely subtle 2. … 3.Fairly Subtle 4. … 5.Obvious	**Margin** Margin definition quality	1.Poorly Defined 2.… 3.… 4.… 5.Sharp
Internal Structure Expected composition	1.Soft Tissue 2.Fluid 3.Fat 4.Air	**Lobulation** Presence of lobular shape	1. Marked 2. … 3. … 4. … 5. None
Calcification Calcification pattern	1.Popcorn 2.Laminated 3.Solid 4.Non-central 5.Central 6.Absent	**Spiculation** Degree of spicules	1. Marked 2. … 3. … 4. … 5. None

Sphericity Roundness	1.Linear 2.… 3.Ovoid 4.… 5.Round	**Texture** Internal density of nodule	1. Non-Solid 2. … 3. Part Solid (mixed) 4. … 5. Solid
Malignancy (Target Variable) Difficulty of detection	1.Highly Unlikely 2.Moderately Unlikely 3.Indeterminate 4.Moderately Suspicious 5.Highly Suspicious		

3.2. Data Pre-processing

While the Yung et al. 2017 study focused on nodules with at least four raters and leveraged a consensus approach for the malignancy classification, we departed from this approach by preprocessing the data as probabilistic vectors to capture disagreement of the raters for each semantic feature. We also made a version of weighted vectors to capture uncertainty of the subjectively rated data.

In preparation, we first cleaned the data by removing 99 nodules with missing gabor statistics, including two nodules that were missing either ratings (NoduleID=2692) or image features (NoduleID=2691). We also removed one nodule that included an undefined rating for Internal Structure = 5 (NoduleID = 1363).

3.2.1.Quantifying disagreement

One of the challenges in the LIDC dataset is related to disagreement, which occurs when two or more raters choose different semantic scores for the same nodule. To capture disagreement of the ratings, we created probabilistic vectors to capture the proportions of agreement between the radiologists participating (Figure 1). To calculate these vectors, we first create a semantic vector whose length corresponds to the number of classes and apply the rater count per class to the corresponding class position. We then divide the vector by the total number of raters participating. In essence, this captures the ratio of raters per class. We refer to these as the "unweighted" probabilistic vectors.

Figure 1. Unweighted probabilistic vectors

Semantic Rater Scores					Raters Participating	Semantic Vector	Probabilistic Semantic Vector
Rater	1	2	3	4	Count	Class Labels: [1 2 3 4 5]	Class Labels: [1 2 3 4 5]
Nodule 1	1	0	0	0	1	[1 0 0 0 0]	[1 0 0 0 0]
Nodule 2	1	1	0	0	2	[2 0 0 0 0]	[1 0 0 0 0]
Nodule 3	1	2	0	0	2	[1 1 0 0 0]	[0.5 0.5 0 0 0]
Nodule 4	1	1	2	0	3	[2 1 0 0 0]	[0.67 0.33 0 0 0]
Nodule 5	1	1	2	2	4	[2 2 0 0 0]	[0.5 0.5 0 0 0]
Nodule 6	1	1	1	2	4	[3 1 0 0 0]	[0.75 0.25 0 0 0]
Nodule 7	1	1	1	1	4	[4 0 0 0 0]	[1 0 0 0 0]
						rater count per class	*agreement* *class count / raters participating*

3.2.2. Quantifying uncertainty

These unweighted vectors do not, however, capture the uncertainty arising from the subjective ratings or variability of the number of raters participating. We address this problem by weighting the

unweighted probabilistic vectors by multiplying the proportion of class count over total potential raters – in this case it is a constant of four. Therefore, nodules with more raters participating would be weighted more heavily than those with fewer raters.

Figure 2. Weighted probabilistic vectors

Semantic Rater Scores					Raters Participating	Semantic Vector						Weighted Probabilistic Semantic Vector					
Rater	1	2	3	4	Count	Class Labels:	1	2	3	4	5	Class Labels:	1	2	3	4	5
Nodule 1	1	0	0	0	1		1	0	0	0	0		0.25	0	0	0	0
Nodule 2	1	1	0	0	2		2	0	0	0	0		0.5	0	0	0	0
Nodule 3	1	2	0	0	2		1	1	0	0	0		0.13	0.13	0	0	0
Nodule 4	1	1	2	0	3		2	1	0	0	0		0.33	0.08	0	0	0
Nodule 5	1	1	2	2	4		2	2	0	0	0		0.25	0.25	0	0	0
Nodule 6	1	1	1	2	4		3	1	0	0	0		0.56	0.06	0	0	0
Nodule 7	1	1	1	1	4		4	0	0	0	0		1	0	0	0	0

$$\left(\frac{\text{Class Counts}}{\text{Raters Participating}}\right) \cdot \left(\frac{\text{Class Count}}{\text{Possible Raters}}\right)$$

agreement certainty

All term calculations are as follows:

- Raters Participating = count of raters participating
- Agreement = class count / raters participating
- Possible Raters = total of potential raters participating (constant: 4)
- Certainty = class count / possible raters
- Weighted Vector = agreement × certainty

3.2.3. The impact of weighting

To illustrate how weighting affects the vectors, we've included a comparison of example vectors below in Figure 3. This comparison illustrates the impact of applying uncertainty weights to the probabilistic vectors. Nodule 1,2, and 7 have different counts of raters selecting a score of 1. The probabilistic vector all show 100% for class 1 for each of these nodules but we want to weight those we are more certain about more. When we take into the effect of *certainty*, we can see that class 1 in the weighted probabilistic vector, nodule 1 with 1 rater shows 25%, nodule 2 with 2 raters, shows 50%, and nodule 7 with 4 raters shows 100%.

Figure 3. Comparison of unweighted and weighted probabilistic vectors

Semantic Rater Scores					Probabilistic Semantic Vector							Weighted Probabilistic Semantic Vector						
Rater	1	2	3	4	Class Labels:	1	2	3	4	5		Class Labels:	1	2	3	4	5	
Nodule 1	1	0	0	0	1 rater	1	0	0	0	0			0.25	0	0	0	0	
Nodule 2	1	1	0	0	2 raters	1	0	0	0	0			0.5	0	0	0	0	
Nodule 3	1	2	0	0		0.5	0.5	0	0	0			0.13	0.13	0	0	0	
Nodule 4	1	1	2	0		0.67	0.33	0	0	0			0.33	0.08	0	0	0	
Nodule 5	1	1	2	2		0.5	0.5	0	0	0			0.25	0.25	0	0	0	
Nodule 6	1	1	1	2		0.75	0.25	0	0	0			0.56	0.06	0	0	0	
Nodule 7	1	1	1	1	4 raters	1	0	0	0	0			1	0	0	0	0	

agreement agreement x certainty

3.3. K-Nearest Neighbors (KNN)

The KNN algorithm first proposed by (Fix and Hodges 1951) is a non-parametric, lazy learning

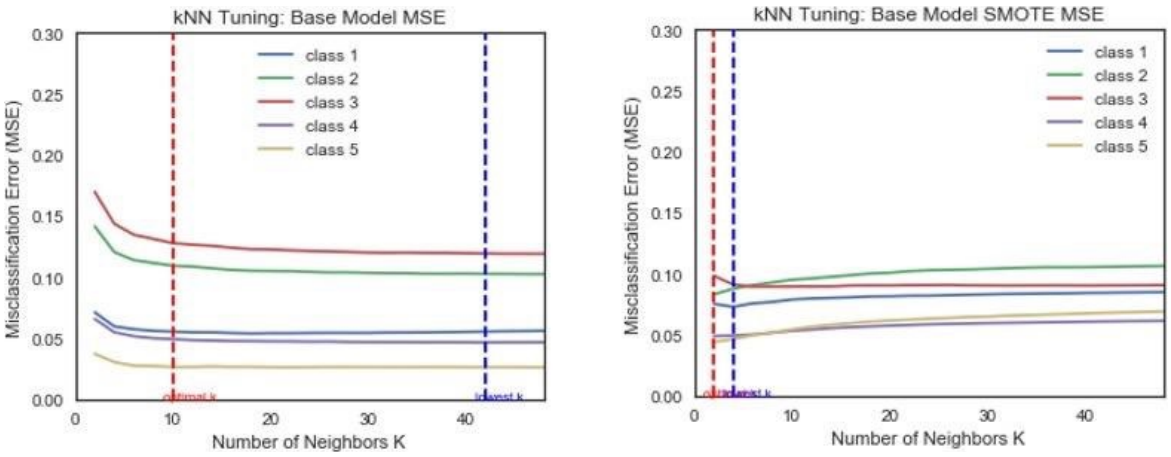
algorithm that can be used for regression and classification problems. For each instance that needs to be classified, KNN algorithm matches it up with its k closest neighbors based on a similarity function. The instance's label is determined based on the majority class of its neighbors. Since we are predicting vector of probabilities for each instance, we first identified the nearest neighbors for each instance then compute the mean of each class of the neighboring instances.

K-fold cross validation is applied to the model to increase its robustness. K-fold cross validation works by partitioning the original dataset by equal-sized subsets (or *k* folds) in which one of the subsets is retained as the testing set and the remaining ones are used as the training data. The model is fit on the training set and evaluated on the test set. This process is repeated *k* times and the error, in our case mean squared error, from the trials is averaged. The advantage of this validation method is that every data point is included in a test set exactly once and in the training set *k-1* times.

We also tuned the model parameter for optimal number of neighbors by assessing the vector average of the averaged mean squared error (MSE) per feature class for each value of *k* neighbors. MSEs tend to plateau when increasing *k* values, so improvement of the error becomes minimal. To choose an optimal *k* value, we implemented an "early stopping" mechanism which initiates when the difference between current and previous MSEs reach a threshold that we defined as 0.1. That is, if MSE*k* - MSE*k-1* < 0.1, then *k-1* was chosen as the "optimal" *k*. Figure 4 includes two example plots showing the difference in optimal *k* vs lowest MSE *k* for the base model with and without applying SMOTE.

To run the KNN models, we used the Sci-kit Learn NearestNeighbors module for Python (Pedregosa et al. 2011). We used the same default parameters for each subsequent model - algorithm='auto', metric= 'minkowski', and leaves=30.

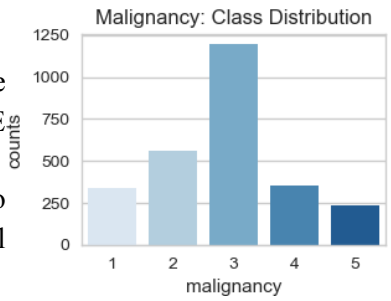
Figure 4. Tuning k (red = optimal k, blue = lowest k)



3.4. Balancing Data (SMOTE)

In our preliminary data analysis, we confirmed the finding of Yung et al. 2017 that the malignancy feature was unbalanced using the consensus – or mode and ceiling of the mean rating. As a result, the performance of the models can be heavily biased toward the majority class, which is why we consider applying the SMOTE algorithm. See figure 5.

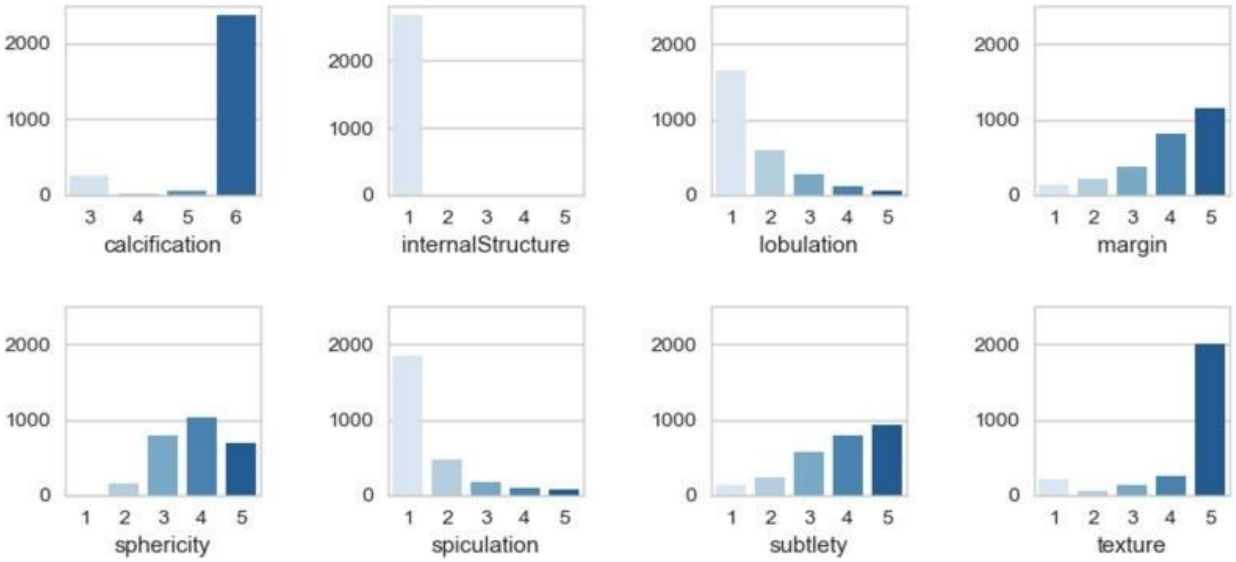
Figure 5. Malignancy class distribution



We also found the remaining semantic variables were also unbalanced, although we only chose to balance malignancy. These will

come into play again when we discuss impact of weighting on the semantic features.

Figure 6. Summary of Semantic Class Distributions



Generally, sampling can be achieved by under-sampling the majority class or over-sampling the minority class. But each technique has its disadvantages. While random under sampling suffers from the loss of potentially useful information, random over sampling suffers from the overfitting issue. Based on results obtained by Yung et al. 2017 for evaluating algorithms addressing class imbalances in the LIDC data set, we chose to apply the Synthetic Minority Over-sampling Technique (SMOTE), which synthetically generates new minority samples from its K-nearest neighbors (Chawla et al. 2002). We used the Sci-kit Learn SMOTE module for Python (Pedregosa et al. 2011) to perform this over-sampling technique.

3.5. Experimental Design

Our experimental design is comprised of iterations of models with increasing complexity. We address each disagreement and uncertainty, then combinations of both using a customized neighbors-based regression algorithm. The most complex model leverages a system of cascading classifiers.

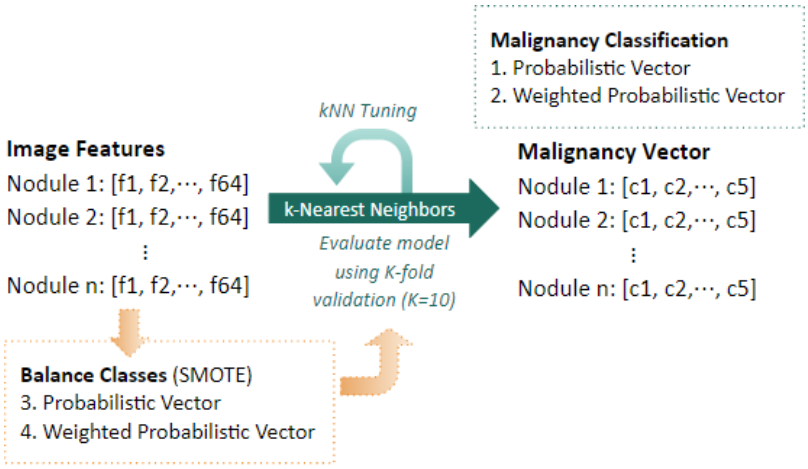
3.5.1.Base Models

The base model predicts the probabilistic malignancy vector directly from image feature inputs using neighbors-based regression. We run the model first on the unweighted data set to address disagreement, and then again with the weighted data to address uncertainty. See Figure 7.

3.5.2. Cascading Classifier Models

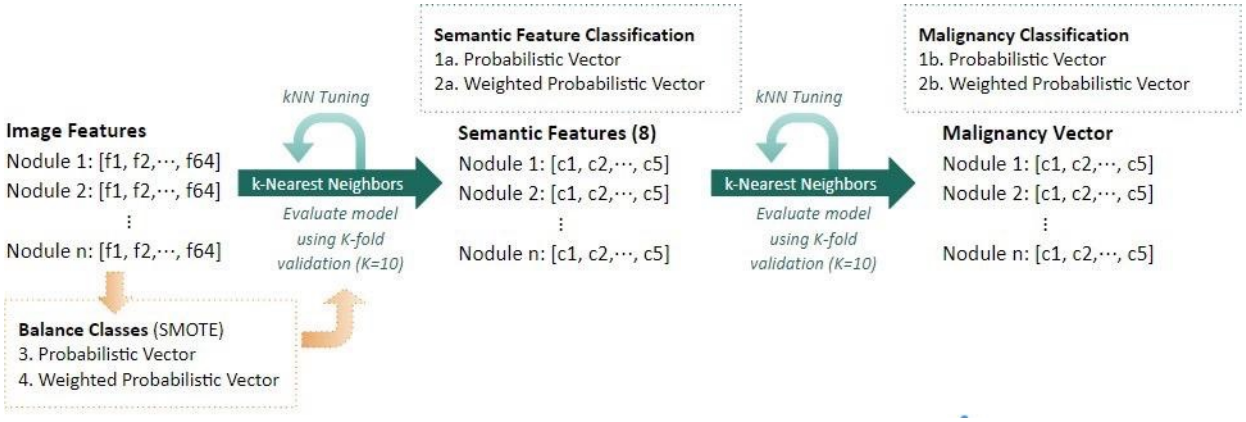
A cascading classifier takes the output predictions from one classifier to use as inputs for the next classifier in the cascade. (Gama and Brazdil 2000). We use the cascading classifier model to further embrace uncertainty by using the measured image features to predict the probabilistic semantic vectors independently. We hypothesize that by doing this we will further increase certainty of the subject semantic ratings. The resulting predicted semantic probabilistic

Figure 7. Base model design



vectors are then fed as inputs to the second cascade to predict the probabilistic malignancy vectors. We applied the cascading classifier to both unweighted and weighted vectors. See Figure 8.

Figure 8. Cascading classifier model design



3.5.3. Models with Balanced Data

For the final stage of the experiment, we run additional iterations on both the base model and cascading classifier model applying SMOTE to balance the data against the probabilistic malignancy vector using the malignancy class mode or ceil of the mean. See Figures 7 and 8.

3.6. Evaluation Metrics

To evaluate and compare the final eight models, we used two evaluation metrics methods – the averaged mean-squared error (MSE) of the vector and Area Under the Distance Threshold Curve (AUCdt) using Jeffreys Divergence as the distance measure. We chose MSE for the continuous nature of the class probabilities, and we chose AUCdt for its direct applicability to probabilistic, multi-class problems (Zinovev et al. 2011b).

3.6.1. Mean Squared Error (MSE)

To calculate the MSE per class (*MSEclass*), we averaged the resulting MSE scores from our 10-fold cross validation method for each class probability in the vector. See Equation 1.

Equation 1. MSE of the probabilistic class

$$MSE_{class} = \frac{1}{K}\sum_{i=1}^{K}(Y_i - \hat{Y}_i)^2 \text{, where } K = folds\ in\ K fold\ validation$$

To calculate the overall model MSE (MSE_{model}) that we used to compare different models, we averaged the vector of MSE_{class} scores. See Equation 2.

Equation 2. MSE of the overall model

$$MSE_{model} = \frac{1}{n}\sum_{i=1}^{c} MSE_{class_c} \text{, where } n = number\ of\ classes, c = malignancy\ class$$

3.6.2. Area Under the Distance Threshold Curve (AUCdt)

AUCdt was introduced by Zinovev et al. 2011b where the authors defined the distance curve as follows:

> *"Let y be a sequence of instance labels, y=[y_1,y_2...y_j...$y_{|S|}$] where |S| is the number of instances and each y is a discrete probability density function over the label set. Also, H(I) be a sequence of predicted labels, P=[H(I_1), H(I_2),...H(I_j)...H($I_{|S|}$)] where each H(I) is discrete probability density function over the label set."*

They defined the distance-threshold curve as follows:

Equation 3. Distance-threshold curve

$$\frac{\sum_{j=1}^{N}[\text{Dist}\,(y_j, H(I_j)) \leq x]}{N}$$

And they defined the area under the distance-threshold curve as follows:

Equation 4. Area under the distance-threshold curve

$$\int_0^1 \frac{\sum_{j=1}^{|S|}[Dist(y_j, H(I_j)) \leq x]}{|S|}\, dx$$

The Jeffrey Divergence, which can be used for vector distance, is substituted as *d* above. It is calculated as follows:

Equation 5. Jeffrey Divergence

$$d_{JD}(A,B) = \sum_{i=1}^{n}[A_i \log\left(\frac{A_i}{\frac{A_i+B_i}{2}}\right) + B_i \log\left(\frac{B_i}{\frac{A_i+B_i}{2}}\right)]$$

To generate the curve, we use Jeffrey Divergence to compute distance between the ratings probability vectors and predicted probability vectors. Then we used different distance thresholds between 0 and 1. As we increased the distance threshold, an increasing number of elements become included within the threshold, and we must consider different thresholds for the distance distributions for the classification to be considered accurate. The area under the distance thresholds curve was used as metric to compare the accuracy of different models.

4. RESULTS

In this section, the results of the proposed approaches are presented using the two validation metrics. We used 10-fold cross validation throughout all our models and k-tuning steps. The predictive performance of our base model using unweighted test data (capturing disagreement) realized an MSE score of 0.0854. By quantifying both uncertainty and disagreement in our cascade model we realized an MSE score of 0.0200, a 77% decrease – or improvement – in this evaluation metric (Table 3). Likewise, our AUCdt score of our base model on test data was 0.4833 (equivalent to a random result), while our cascade model resulted in an AUCdt value of 0.7630, a 58% improvement (Table 3).

With the exception of the Base Model with SMOTE and the Cascade Model with SMOTE, the difference between test and train results fell within a ±10% threshold, so overfitting or underfitting was not an insurmountable issue.

Table 3. Comparison of final model results

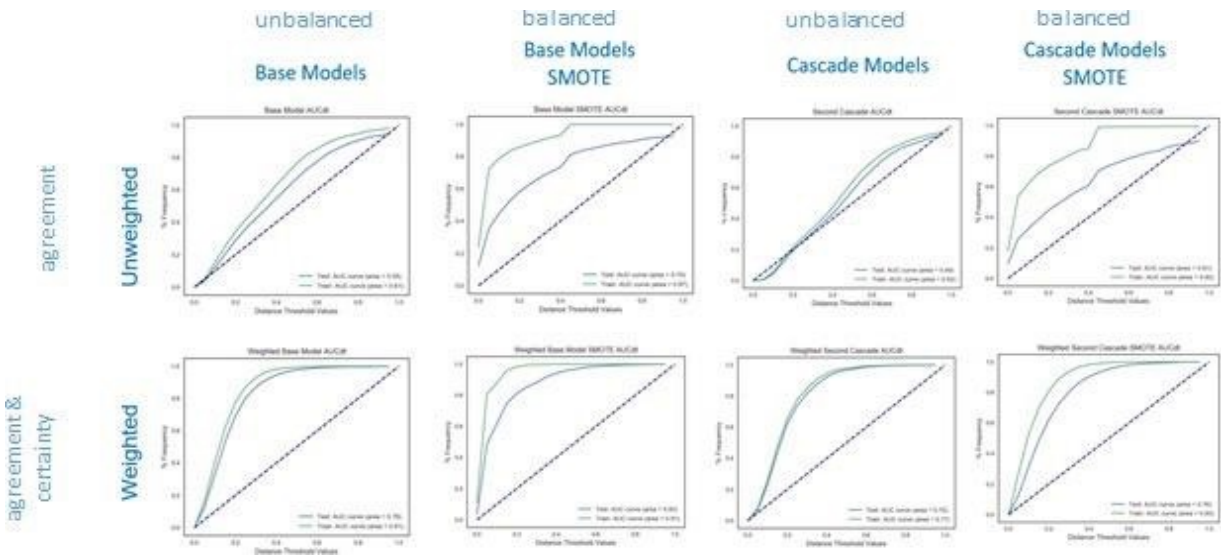
Comparison of all models	**MSE lower is better**		**AUCdt higher is better**	
Models	**Train**	**Test**	**Train**	**Test**
Base Model	0.0690	0.0854	0.5477	0.4833
Cascade Model	0.0730	0.0805	0.5204	0.4912
Base Model with SMOTE	0.0195	0.0751	0.8393	0.6287
Cascade Model SMOTE	0.0225	0.0719	0.8214	0.6148
Base Model Weighted	0.0173	0.0214	0.7769	0.7468
Cascade Model Weighted	0.0184	0.0200	0.7684	0.7542
Cascade Model Weighted SMOTE	0.0124	0.0200	0.8309	0.7630
Base Model Weighted with SMOTE	0.0046	0.0175	0.8955	0.8001

Figure 9 shows AUCdt curves for each model for both unweighted, weighted, and balanced data. When comparing across the unweighted AUCdt results (top row), we can see that balancing data with SMOTE improves the model predictions – that is, the area under the curve is increased. The over- and under-fitted results of the unweighted, balanced data is noticeable in the larger gap between the test and train curve. For the other models, we see that differences do not exceed ±10%. When comparing the unweighted and weighted results, we can generally see that weighting the malignancy vectors to account for disagreement and uncertainty provided more lift than not applying a weighting factor.

4.1 Validation Results

We performed a validation run of 100 trials of both the base and cascade models. Results of this run were aligned with, albeit slightly less favorable than, evaluation values for our initial test runs. We continue to see the trade-off of between class balance and model overfitting using SMOTE on the Malignancy class, as you can see based on the differences between training and testing results in the balanced plots in Figure 9.

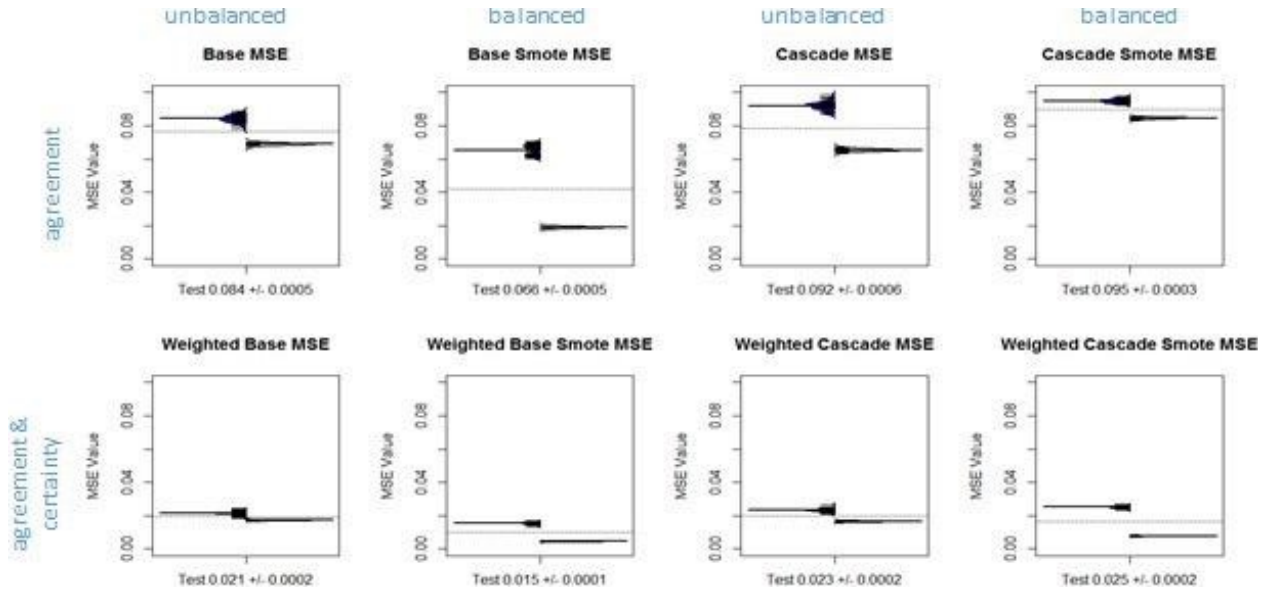
Figure 9. AUCdt curves for all final models



4.1.1. Model Evaluation: MSE with confidence intervals

We reviewed MSE results for the validation run of 100 trials. Each bean plot in Figure 10 shows the distribution and mean of MSE values for both training set (on the right, in grey) and testing set (on the left, in blue). Note that MSE values for the testing set are higher and wider dispersed than those for training set. You can also clearly see the dramatically lower MSE scores resulting from our weighted vector approach on the bottom row of plots.

Figure 10. Model evaluation. MSE with confidence intervals



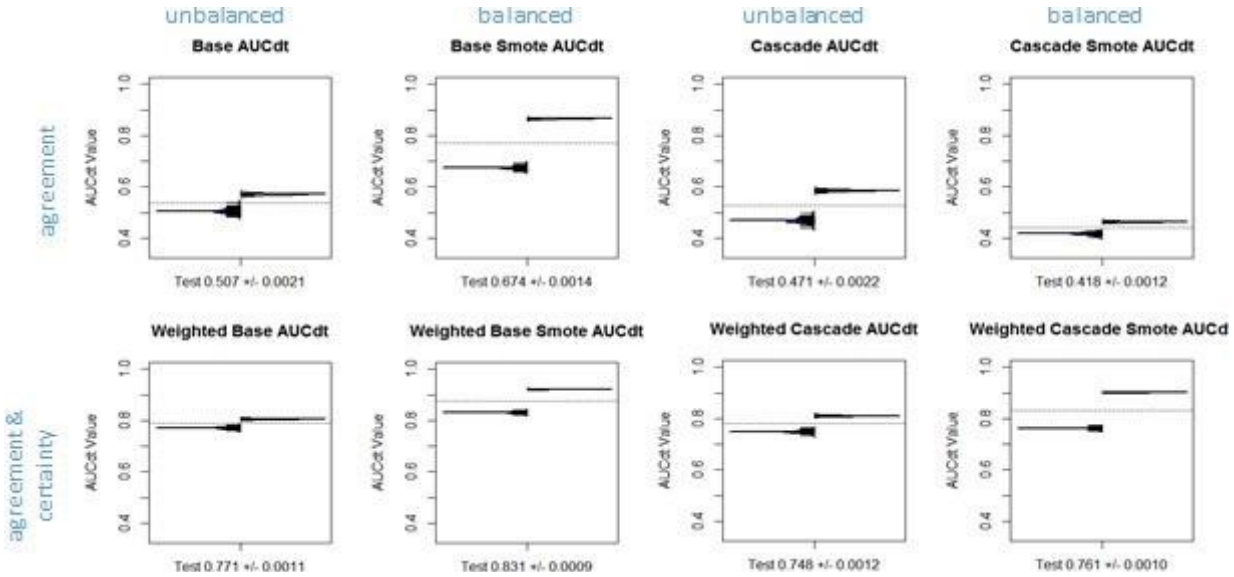
Left-side (test data) Right-side (train data)

4.1.2. Model Evaluation: AUCdt with confidence intervals

We also reviewed AUCdt results for the validation run of 100 trials. Each bean plot (Figure 11) shows the distribution of AUCdt values for both training set (on the right) and testing set (on the left). As expected, AUCdt values for test data are lower and wider dispersed than those for train data. You can also clearly see the dramatically higher AUCdt values resulting from the weighted vector approach on the

bottom row of plots.

Figure 11.Model evaluation. AUC with confidence intervals



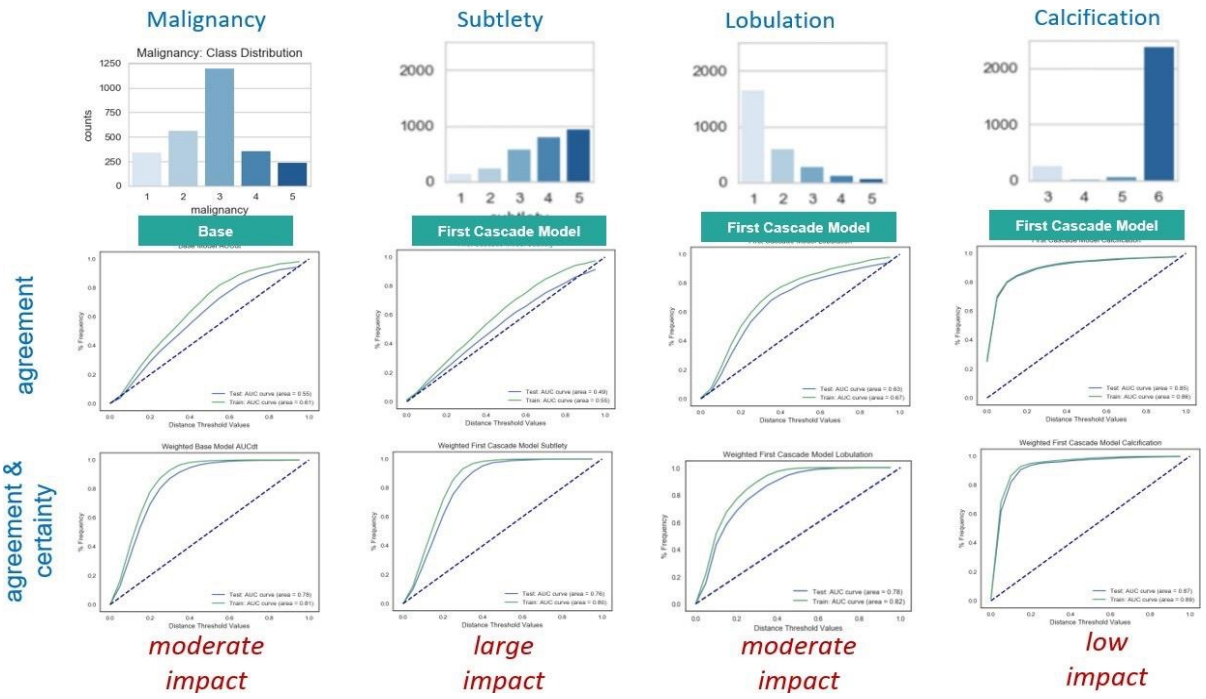
Left-side (test data) Right-side (train data)

5. DISCUSSION

5.1. Impact of capturing Uncertainty in Semantic Features

Impact of weighting the vectors for certainty varied across semantic vectors. Figure 12, which shows the nodule distribution and AUCdt curve by semantic features, helps visualize the impact. We saw that high class-imbalanced data, such as calcification, was barely impacted by the weighting as the model predictions were generally biased towards majority class. Balanced semantic features, such as Subtlety, saw much larger improvement, as we add more "certainty" to the prediction. Malignancy and Lobulation,

Figure 12. Nodule Distribution and AUCdt curves (blue = test, green = train)



which were imbalanced, but not to the degree as Calcification and Internal Structure, showed moderate impact. And because of imbalance issue, the MSE values of calcification and internal structure also show a different trend than others.

5.2. Impact of SMOTE

From the results section we saw that the differences between training and testing sets in the unweighted models with SMOTE are larger than 10%, which indicates an overfitting problem. We speculate that the reason why the unweighted models with SMOTE display an overfitting problem is because we apply SMOTE to the full dataset before feeding it into our algorithm, as opposed to applying it to the subset of data during cross-validation. Although SMOTE effectively forces the decision region of the minority class to become more general, partially solving the generalization problem, it has nothing to do with cross-validation. Because if the nearest neighbors of minority class observations in the training set end up in the testing set, their information is partially captured by the synthetic data in the training set, which is defeating the purpose of cross-validation.

5.3. Computability between Base and Cascade

In terms of computability, the cascade model took much longer compared the base model by 5 to 21 times due to the complexity of the model, which includes training a model for each of the eight semantic features from the image features and then training a malignancy model from the semantic features. Table 4 shows the time it took to run the Base and Cascade models with the *k* neighbors tuning and without the Tuning for the MSE and AUCdt metric. The non-parametric, non-linear KNN is already a rather computationally expensive model; adding the additional cascading step and increasing data size may cause additional computational issues.

Table 4. Model computability

	Base Model				**Cascade Model**			
	Model	**Model with SMOTE**	**Weighted Model**	**Weighted Model with SMOTE**	**Model**	**Model with SMOTE**	**Weighted Model**	**Weighted Model with SMOTE**
with Tuning	00:00:54	00:02:26	00:00:55	00:01:57	00:06:37	00:25:17	00:14:30	00:31:54
with MSE	00:00:06	00:00:04	00:00:04	00:00:05	00:01:10	00:01:01	00:00:54	00:01:46
with AUCdt	00:00:06	00:00:05	00:00:05	00:00:05	00:00:35	00:00:31	00:00:27	00:00:51

5.4. Comparison to previous work

We compared our results to those of another KNN probabilistic vector classifier (Williams and Raicu 2013) which selected only nodules with four expert raters and created the dataset up to 5,000 observations. Our best model was only 1.4% lower even though we chose to include any number of radiologist raters, thereby embracing further uncertainty of the data.

Comparison of AUCdt JD	**5-NN Classifier (Williams and Raicu 2013)**	**Cascade Model Weighted**	**Cascade Model Weighted with SMOTE**	**Base Model Weighted with SMOTE**

Training	0.8620	0.7684	0.8309	0.8955
Testing	0.8141	0.7542	0.7630	0.8001

6. CONCLUSION AND FUTURE WORKS

6.1. Conclusion

In this work we demonstrated an iterative classification process in determining the reliability of CAD systems with both uncertainty and multiple rater disagreement. We developed a two-phase KNN classification model with base model and a cascade model. In the base model, we predicted malignancy using the 64 image features. In the cascading phase, we faced two challenges, uncertainty and agreement. We embraced uncertainty in two ways. First, using the measured image features to predict the probabilistic semantic feature vectors. Second, we resolved the malignancy class imbalance by applying SMOTE to the dataset to balance it. Addressing disagreement was accomplished by first transforming the individual ratings of the experts into probabilistic vectors. We then weighted these vector elements based on the percent of possible experts (4) who rendered an opinion. We saw that addressing uncertainty alone provided little or no improvement in evaluation metrics of the base model, but addressing both uncertainty and disagreement provided significant improvement in evaluation metrics. Our conclusion thus shows the validity of embracing both disagreement and uncertainty within multiple rater data as evidenced by the gains in evaluation metric values and as evidenced by the performance of the weighted cascade model.

6.2. Future Works

The LIDC data is generally high dimensional in nature and while some machine learning techniques such as SVM among can handle high dimensional data, not all techniques can. Our initial experimental design included attempting Principal Component Analysis (PCA) as a feature reduction technique. Therefore, an obvious open question is whether feature reduction can improve performance of the model. Specifically, a future work would involve repeating steps one through four using a smaller feature space data.

A second open question is around scalability. While a great choice for complex tasks such as developing CAD systems, k-NN can be computationally expensive, especially with increasing dimension of data as there isn't any "training" done prior to prediction. In addition, with increased data sets, computational challenges are generally expected. Consequently, as the LIDC data increases in size, an implementation of our approach could involve a parallel distributed system to optimize scalability.

The third open item would be to consider other machine learning techniques. Most literature around multi- label datasets have focused heavily on problem transformation, e.g. transforming a multi-label problem into a binary classification or multi-class classification problem. Recently, algorithm adaptation has taken center stage and we believe more work will center around this approach, especially when one considers the limitations of (and associated information loss of) problem transformation techniques. As such, there are many opportunities for algorithm adaptation classification techniques.

The fourth open question is about generalizable. By oversampling only on the training data, none of the information in the testing data is being used to create synthetic observations. So these results could be generalizable.

6.3. Other data sets

In our study, 9 subjective features, composed of ratings by 4 independent radiologist experts according to their experiences, gave us the biggest challenge. Medical images provide the information used by radiologists to determine a diagnosis of cancer. There are two common procedures depending on the type of visceral organs: CT scans and MRI.

Therefore, we found that the website http://www.cardiacatlas.org/studies/scmr-consensus-data/ provides a dataset, named SCMR consensus contour data, which has a similar issue. This dataset contains cardiovascular MRI studies of 15 patients. These are read by seven independent experts, according to their own standard operating protocols and their preferred tools. Since the dataset has both challenges of uncertainty and multiple experts, we consider that our dataset as well as all medical image datasets used in CADx systems have the same challenge.

Our study focuses on the reliability of CAD with uncertain inputs. Rating by different radiologists is the source of uncertainty. In other domains, if the features of a dataset should be judged by people, our work would be applied to it. We found that a wine-tasting dataset has many factors, which depend on critics' scoring, such as aesthetics, pleasure, complexity, color, appearance, odor, aroma. Also, the website bordoverview.com, contains ratings of different wines assessed by world-renowned critics from the United States and Europe.

ACKNOWLEDGEMENTS

This research task held personal significance for several members of the research team whose family members have succumbed to this disease and therefore appreciate the opportunity to make a contribution.

We also want to acknowledge and thank Matt Yung (Yung et al. 2017) for sharing his dataset and providing guidance on his preprocessing techniques.

REFERENCES

1.American Cancer Society. https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html. Accessed 14 February 2018.

The Cancer Imaging Archive. Available at http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX. Accessed 11 May 2011.

2.Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, van Beek EJR, Yankelevitz D, et al.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. Medical Physics, 38: 915--931, 2011.

Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16: 321–357, 2002.

15.Dhara, A., Mukhopadhyay, S., Dutta, A., Garg, M., and Khandelwal, N.: A Combination of Shape and Texture Features for Classification of Pulmonary Nodules in Lung CT Images. J Digit Imaging, 29(4): 466-475, 2016.

Fix, E., Hodges, J.: Discriminatory analysis, nonparametric discrimination: consistency properties. Technical Report, 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

J. and Brazdil, P.: Cascade Generalization. Machine Learning, 41: 315–343, 2000.

Gibaja, E., & Ventura, S.: A Tutorial on Multi-Label Learning. ACM COMPUT SURV, DOI: 10.1145/2716262, 2015

Hancock, M. C., & Magnan, J. F.: Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods. J Med Imaging, DOI: 10.1117/1.JMI. 3.4.044504, 2016

8.Jarvik JG, Deyo RA.: Moderate versus mediocre: the reliability of spine MR data interpretations. Radiology, 250(1): 15–17, 2019.

12.Kaya, A.; Can, AB.: A Weighted Rule Based Method for Predicting Malignancy of Pulmonary Nodules by Nodules Characteristics. Journal of Biomedical Informatics, 56: 60-79, 2015.

4.Khaw, A., Angermaier, A., Michel, P., Kirsch, M., Kessler, C., Langner, S.: Inter-rater Agreement in Three Perfusion-Computed Tomography Evaluation Methods before Endovascular Therapy for Acute Ischemic Stroke. Journal of stroke and cerebrovascular diseases: the official journal of National Stroke Association, 25(4): 960-8, 2016

Kilem L. Gwet, P.: On The Krippendorff's Alpha Coefficient. Communication Methods and Measures, 2011

3.Kunio Doi, P.: Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. Comput Med Imaging Graph, 31(4-5): 198-211, 2007.

Madjarov, G., Kocev, D., Gjorgjevikj, D., & Dzeroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, 45 (9): 3084-3104, 2012

Moyano J, Gibaja E, Cios K, Ventura S. Review of ensembles of multi-label classifiers: Models, experimental study and prospects. Information Fusion, 44: 33-45, 2018.

9.Opfer, R., & Wiemker, R.: Performance Analysis for Computer Aided Lung Nodule Detection on LIDC Data. Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment. San Diego, California: SPIE, 2007

Pedregosa et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12: 2825-2830, 2011

5.Pérez-Gómez, B., Ruiz, F., Martínez, I., Casals, M., Miranda, J., Sánchez-Contador, C., Vidal, C., Llobet, R., Pollán, M., Salas, D.: Women's features and inter-/intra-rater agreement on mammographic density assessment in full-field digital mammograms (DDM-SPAIN). Breast Cancer Research and

Treatment, 132 (1): 287-295, 2012.

Raicu, D. S., Varutbangkul, E., Furst, J. D., & III, S. G.: Modeling Semantics from Image Data: Opportunities from LIDC. International Journal of Biomedical Engineering and Technology, 3(1/2): 83-113, 2010.

The University of Waikato. Available at https://hdl.handle.net/10289/4645. Accessed 2010.

10.Reeves A.P., Biancardi A.M., Apanasovich T.V., Meyer C.R., MacMahon H., van Beek E.J.R., Kazerooni E.A., Clarke L.P.: The Lung Image Database Consortium (LIDC). A Comparison of Different Size Metrics for Pulmonary Nodule Measurements. *Academic Radiology*, 14 (12): 1475-1485, 2007.

Suinesiaputra, A., Bluemke, D. A., Cowan, B. R., Friedrich, M. G., Kramer, C. M., Kwong, R., . . . Nagel, E.: Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. Journal of Cardiovascular Magnetic Resonance, DOI: 10.1186/s12968-015-0170-9, 28 July 2015

6.Spayne, M., Gard, C., Skelly, J., Miglioretti, D., Vacek, P., Geller, B.: Reproducibility of BI-RADS Breast Density Measures Among Community Radiologists: A Prospective Cohort Study. The Breast Journal, 18(4): 326-333, 2012.

Sydney Williams, M. H. (2013, July). Area under the Distance Threshold Curve as an Evaluation Measure for Probabilistic Classifiers. Machine Learning and Data Mining in Pattern Recognition: 644-657, 2013.

13.Tsoumakas G., Vlahavas I. (2007) Random *k*-Labelsets: An Ensemble Method for Multilabel Classification. Machine Learning: ECML 2007: 406-417, 2007

Valizadegan, H., Nguyen, Q., & Hauskrecht, a. M.: Learning Classification Models from Multiple Experts. J Biomed Inform, DOI: 10.1016/j.jbi.2013.08.007, September 13, 2013.

11.Yung, M., Furst, J., Raicu, D.: Multi-Class Malignancy Prediction with Oversampling Technique Analysis. DePaul University, 2017 (unpublished).

Zapf, A., Castell, S., Morawietz, L., & Karch, A.: Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? BMC Med Res Methodol, DOI: 10.1186/s12874-016-0200-9, August 5, 2016

Zhang M, Zhou Z.: A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8): 1819-1837, August 2014.

14.Zinovev, D., Raicu, D., Furst, J.: Semi-supervised learning approaches for predicting semantic characteristics of lung nodules. Intelligent Decision Technologies, 3(4): 207-217, October 2009.

7.Zinovev, D., Duo, Y., Raicu, D. S., Furst, J., & Armato, S. G.: Consensus Versus Disagreement in Imaging Research: a Case Study Using the LIDC Database. J Digit Imaging, DOI: 10.1007/s10278-011-9445-3, December 23, 2011.

Zinovev, D., Feigenbaum, J., Furst, J., & Raicu, D.: Probabilistic Lung Nodule Classification with Belief Decision Trees. 2011 Annual International Conference of the IEEE EMBS, DOI: 10.1109/IEMBS.2011.6091114, 2011

Zinovev, D., Furst, J., Raicu, D.: Building an ensemble of probabilistic classifiers for lung nodule interpretation. 2011 10th International Conference on ICMLA, DOI:10.1109/ICMLA.2011.44, 2011

APPENDIX

The additional file includes an HTML document of the fully executed code in iPython Notebook format. We did not use a data seed to control the random selection during cross validation, therefore the results vary slightly from those stated in this report. In it you can find the AUCdt evaluation function.