CSC 672 Preliminary Data Analysis

Determining Reliability in CAD Outcomes

Team Name: CADence

Team Leader: Trish Lugtu

Team Members: Yenong Du, Cun Lan, John Misailedes, Collins Nyagaya, Stephanie Wong

Research Question: Can we determine reliability of CAD outcomes?

CSC 672 - Winter 2018

Introduction

The motivation of this study is to determine the reliability of CAD outcomes in the presence of uncertainty in a multi-label classification problem.

The dataset used for the initial exploratory analysis was provided by Yung et al. from a previous study[1]. The data set is accessible at the following URL: https://www.dropbox.com/sh/zxq69reu1ez00xm/AACnbSXLHUg-Dqr6YYVwQl2Ba?dl=0.

This dataset was based on the Lung Image Database Consortium (LIDC) dataset, but has been preprocessed to contain only one slice per nodule with the the maximum intersection of pixels by the rating radiologists with an area greater than 25 pixels. While the final data set used in Yung's study focused on nodules with at least four raters and leveraged a consensus approach for the malignancy classification, we needed to process the data to capture agreement between raters, as well as the number of raters.

Research Question

One common treatment to handle multiple raters is the use of a consensus model. A consensus model is when some form of aggregation is applied to a feature, such as mode or mean. However, nuances of the model can be lost through this aggregation.

To recapture the agility of the model, we need to address two challenges. One challenge is that of disagreement and the other of uncertainty. When multiple raters participate, whether or not their ratings agree introduces an added complexity that is lost in a consensus model. The second challenge is to overcome the uncertainty of the data. We operate on the assumption that more raters participating will orchestrate less uncertain data. That is, the likelihood of reaching a "ground truth" is higher when more raters participate. Conversely, the likelihood of reaching a "ground truth" with only one rater participating will be at its lowest.

To address these challenges, we will capture new features - the number of raters participating and the number of raters in agreement - then we will attempt to build an ensemble models using probabilistic vector classification to predict this multi-class problem. The ensemble models will include KNN, BDT, SVM, and as a stretch goal, ANN.

Therefore, our research question is to determine whether we can improve classification performance - accuracy, sensitivity, and AUC - of computer aided diagnosis (CAD) outcomes by embracing both uncertainty and disagreement in our model using probabilistic vector classification in ensemble models.

Data Description

Our dataset contains 2686 nodule slices with semantic features (9), size features (7), shape features(8), intensity features (9), and texture features (40) totalling nine semantic features (table 1) and 64 image features (table 2). There were two nodules in the original data that were missing either ratings (NoduleID=2692) or image features (NoduleID=2691), which were subsequently dropped. Our target variable is malignancy, a multi-class label which we will treat as a probabilistic vector based on the multiple raters in our final methodology. Descriptions of the data are included below.

Semantic Features (Radiologist-rated)

Table 1. Semantic Features[2]

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

Texture	Internal density of nodule	1. Non-Solid 2. ... 3. Part Solid (mixed) 4. ... 5. Solid	Ordinal
Malignancy (Target Variable)	Likelihood of malignancy	1. Highly Unlikely 2. Moderately Unlikely 3. Indeterminate 4. Moderately Suspicious 5. Highly Suspicious	Ordinal

Image Features

Table 2. Image Features[2]

Category	Feature/Description	Data Type
Size	• Area • Convex Area • Perimeter • Convex Perimeter • Equivalent Diameter • Major Axis Length • Minor Axis Length	Continuous
2D Shape	• Circularity • Roughness • Elongation • Compactness • Eccentricity • Solidity • Extent • Radial Distance SD	Continuous
Intensity	• Minimum Intensity • Maximum Intensity • Mean Intensity • Intensity SD • Minimum Intensity Background • Maximum Intensity Background • Mean Intensity Background • SD Intensity Background • Intensity Difference	Continuous
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	

Exploratory Data Analysis: Semantic Features

For the initial analysis of features with multiple ratings, we employed a consensus approach. For each feature - malignancy, subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, and texture - we either took the mode of the rating or the ceiling of the mean in the mode's absence. This allowed us to get an overall sense of the shape of our data without inflating nodule instances when multiple raters were participating.

We also added a "participating radiologist"count (NumRads), as well as an "agreement" count of mode malignancy scores when available (NumAgree) (fig. 1).

Figure 1. Rater Count and Agreement per Nodule

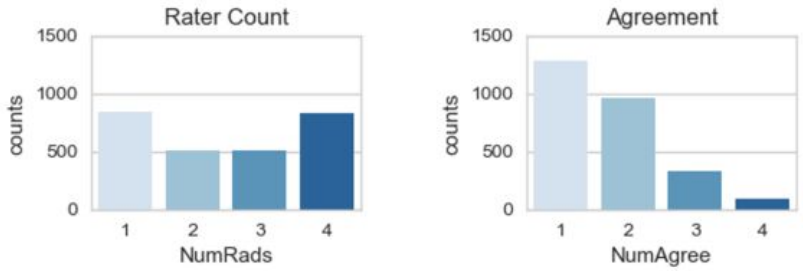


Figure 2. Rater Count, Agreement, and Their Effects on Malignancy

Malignancy Class Distribution

The class distribution of malignancy (fig. 3, tbl. 3) shows the imbalance between the malignancy classes. The peak of the distribution represents malignancy class 3, which means the nodule malignancy is indeterminate in nature. An indeterminate classification may lead to other recommendations, including undergoing advanced medical procedure such as a biopsy or the physician may choose the "wait-and-see" approach. Our approach to addressing the imbalance issue is to apply SMOTE to create synthetic features of the less populated classes.

Figure 3. Malignancy Class Distribution

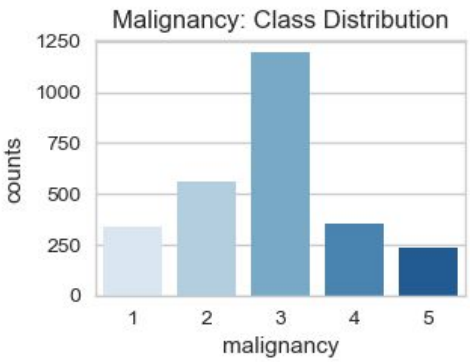


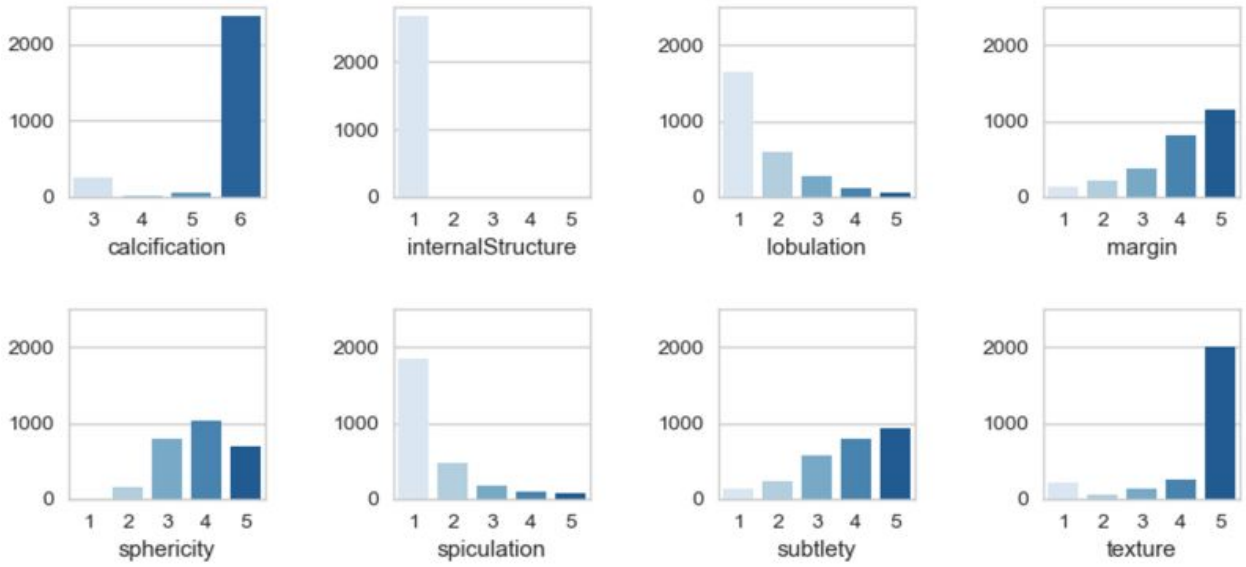
Table 3. Malignancy Class Proportions

malignancy class	count	percent
1	337	12.6%
2	561	20.9%
3	1196	44.5%
4	357	13.3%
5	235	8.8%

Semantic Features

The distributions (fig. 4) of each semantic feature are below. Internal structure and calcification - the categorical features - indicate a majority of nodules as having a soft tissue structure and absence of calcification, correspondingly. Nodules tend to be more obvious in their definition (subtlety, margin, texture), ovoid to round (sphericity), and more marked in their lobulation and spiculation

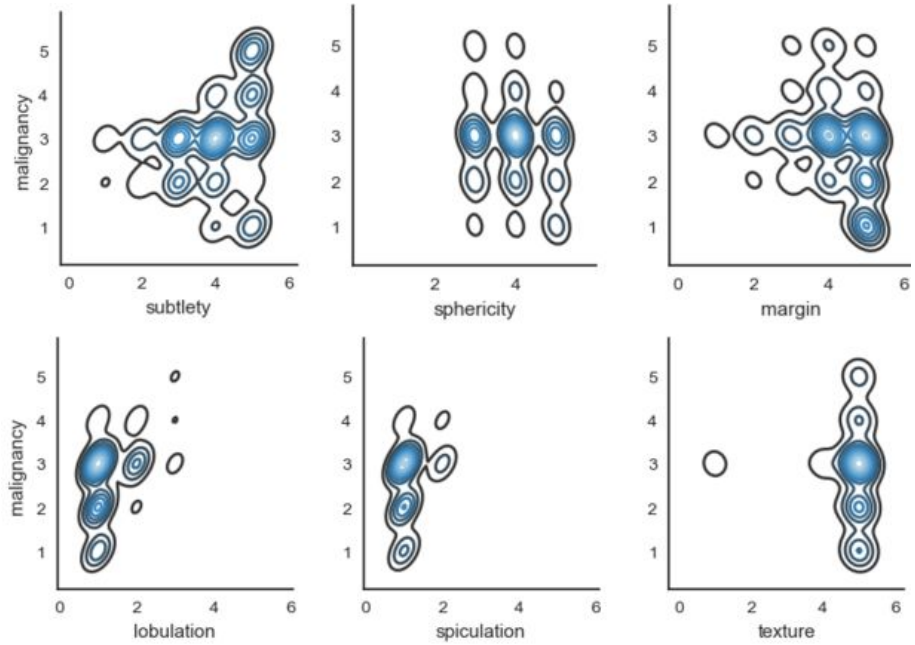
Figure 4. Semantic Feature Distributions



Relationships between Malignancy and Ordinal Semantic Features

The bivariate kernel density estimation plots (fig. 5) below helps us to understand the general relationships across all malignancy classes and the other eight semantic features. While malignancy classes 4 and 5 are associated with higher spiculation and lobulation, benign nodules (class 1 and 2) are more associated with calcification. However, because the indeterminate class of malignancy holds the highest proportion, the plots in figure 5 show the densest clusters centering over malignancy class 3.

Figure 5. Relationships between Malignancy and Ordinal Semantic Features



Distributions of Ordinal Semantic Features by Malignancy Class

The box plots below (fig. 6) show ordinal semantic feature distributions by malignancy class. The two plots on the left represent benign classes, and the two plots on the right represent malignant classes. This visualization helps to make a couple relationships apparent. For example, malignant nodules are much more marked in both spiculation and lobulation, while each is absent in benign nodules (although outliers exist). Less defined margins seem to also be indicative of malignancy, while rounder sphericity seems to indicate benign nodules.

Figure 6. Distribution of Ordinal Semantic Features by Determinate Malignancy Classes. Benign [1,2] and Malignant [4,5]

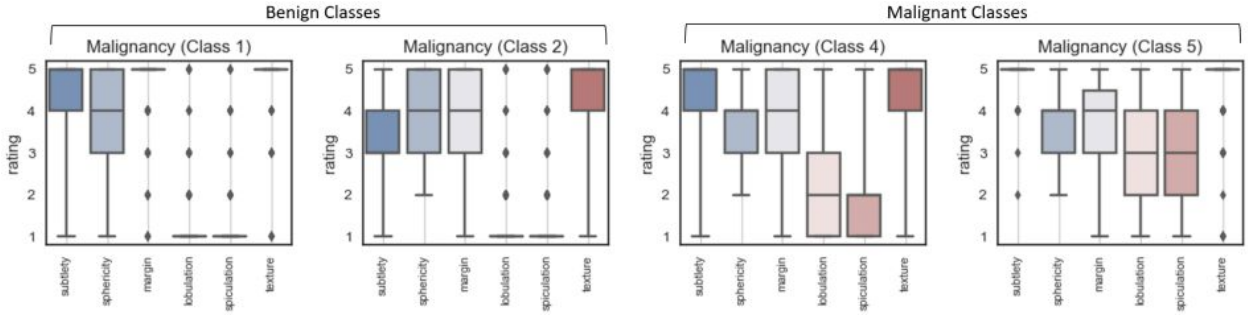


Figure 7 shows the distribution of semantic features when the nodule is deemed indeterminate. It shows characteristics of both malignant and benign nodules.

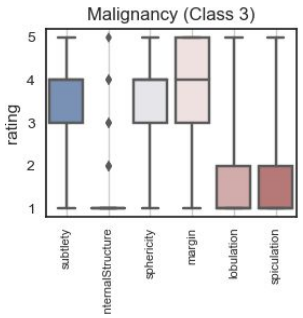
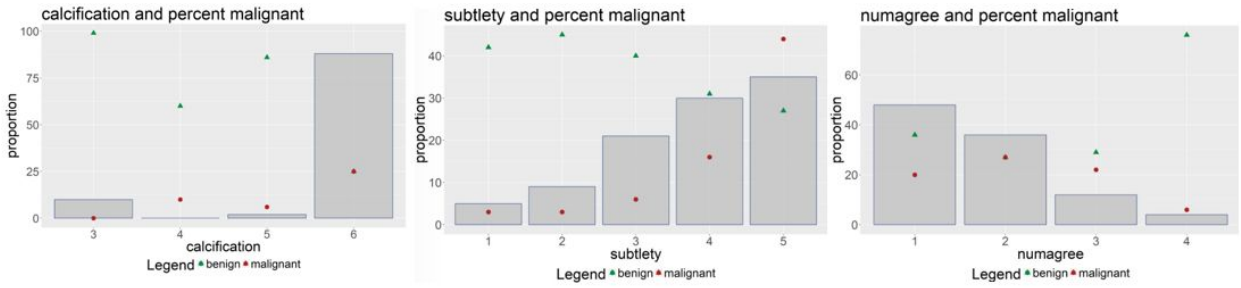


Figure 7. Distribution of Semantic Features by Indeterminate Malignancy Class: Indeterminate [3]

Other Observations

Semantic features, rater agreement and number raters were reviewed against percent malignancy. The grey bars of the plots show the distribution of instances per semantic feature. The red dots show percent malignant and the green triangles show percent benign within each category (Fig.8).

Figure 8. Distribution of features against % malignancy within each group. Benign [1,2] and Malignant [4,5]



For calcification, most nodules fall in the absent category 6. Compared to the other calcification categories, this segment shows the highest % malignancy. A large portion of these nodules fall within the indeterminate category. Segment 3 (solid), shows the second highest distribution; the % benign very high here. In terms of subtlety, nodules that were obvious showed a higher % malignancy while those that were very subtle showed a high % benign. When looking the number agreement feature, the %benign is relatively high when all 4 radiologists agreed.

Feature Importance (XGBoost Model)

Semantic, rater count, and agreement features were fed into an XGBoost model (tree based model) simply to see which features were of most important in determining malignancy (1-5). The parameters are max tree depth = 6, step size shrinkage =1, number rounds of boosting = 10, objective = multi: softmax (to classify 5 classes). Dummy variables were created for internal structure and calcification as these are categorical while the ordinal values were treated as numeric.

Based off these variable importance results (fig. 9), we can see that calcification3 (Solid calcification) makes up ~22% and has a negative correlation with malignancy while calcification6 (absent calcification) makes up 11% and has a positive correlation with malignancy. Subtlety and spiculation also are important features. The rater count made up 10% of the importance.

Figure 9. Importance Features and Correlation with Malignancy

Feature	Gain	Correlation
calcification3	0.22	-0.54
subtlety	0.16	0.24
calcification6	0.11	0.55
spiculation	0.10	0.41
numrads	0.10	0.17
lobulation	0.08	0.39
numagree	0.07	-0.07
margin	0.06	-0.21
sphericity	0.05	-0.13
texture	0.03	-0.06
calcification5	0.01	-0.12
internalstructure1	0.00	-0.03
calcification4	0.00	-0.04
internalstructure3	0.00	0.04
internalstructure4	0.00	0.02

Feature Importance (Random Forest Model)

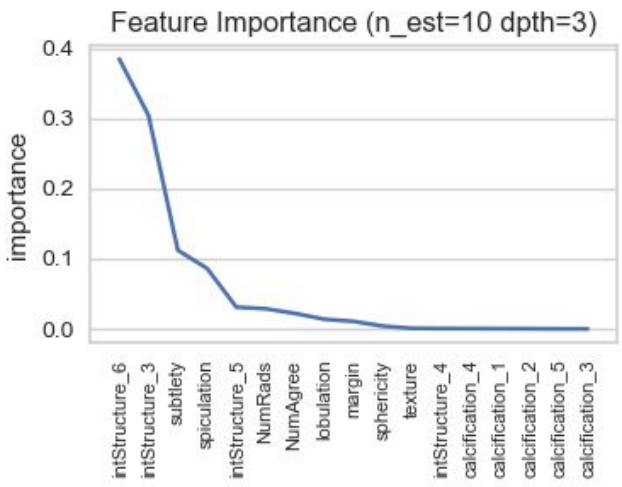
We also tried several Random Forest models with replacement with various depths and enumerations to compare the same features. The model parameters used for the results below were max tree depth = 3 and enumerations=10. The other combinations of depths (3 to 10) and enumerations (10 to 250) yielded similar results in the top ten feature importance, although importance value and order changed slightly between the models. In every case, importance of the features NumRads and NumAgree will prove to be an important addition to our classifiers.

Figure 10a. Top 10 Features

Feature ranking:

	0	1
0	0.3847	intStructure_6
1	0.3041	intStructure_3
2	0.1121	subtlety
3	0.0861	spiculation
4	0.0310	intStructure_5
5	0.0287	NumRads
6	0.0221	NumAgree
7	0.0138	lobulation
8	0.0107	margin
9	0.0040	sphericity
10	0.0009	texture

Figure 10b. Feature Importance Using Random Forest

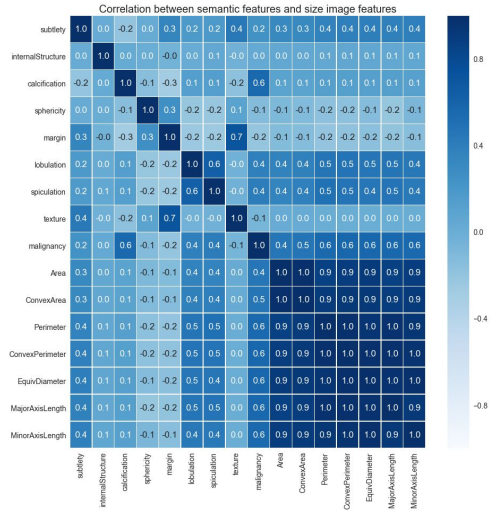


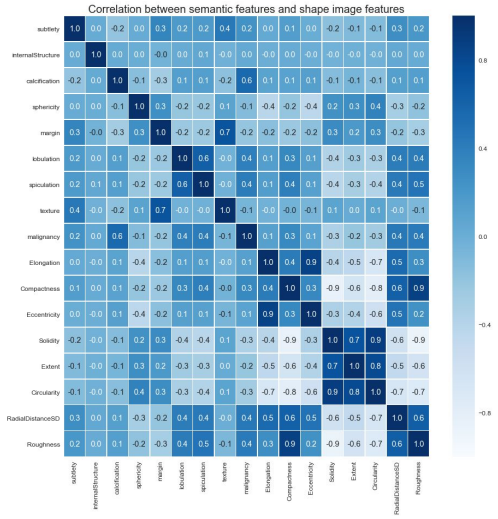
Correlation between Semantic Features

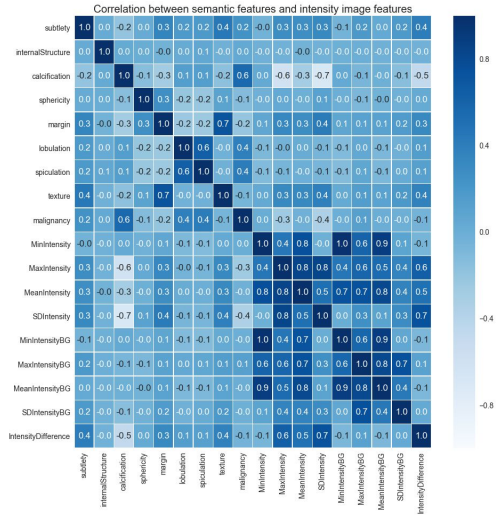
We decided that correlation analysis for the semantic features was unnecessary based on previous research, stating “Furthermore, previous work conducted in our lab has shown that the correlations between different semantic characteristics across different nodules were, in fact, very low.”[3]

Exploratory Data Analysis: Image Features

Correlation between Features







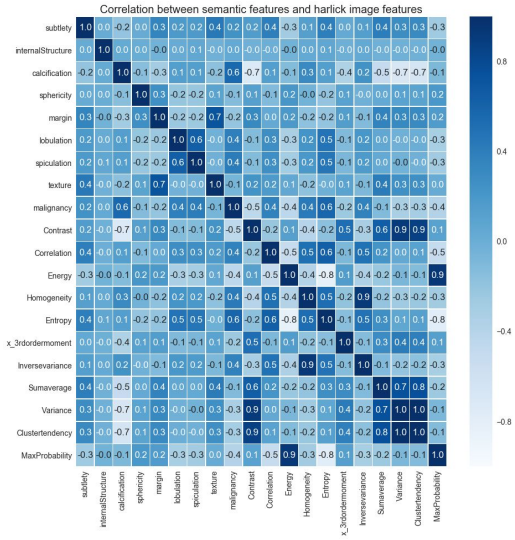


Figure 11a. Correlation between semantic features and image features

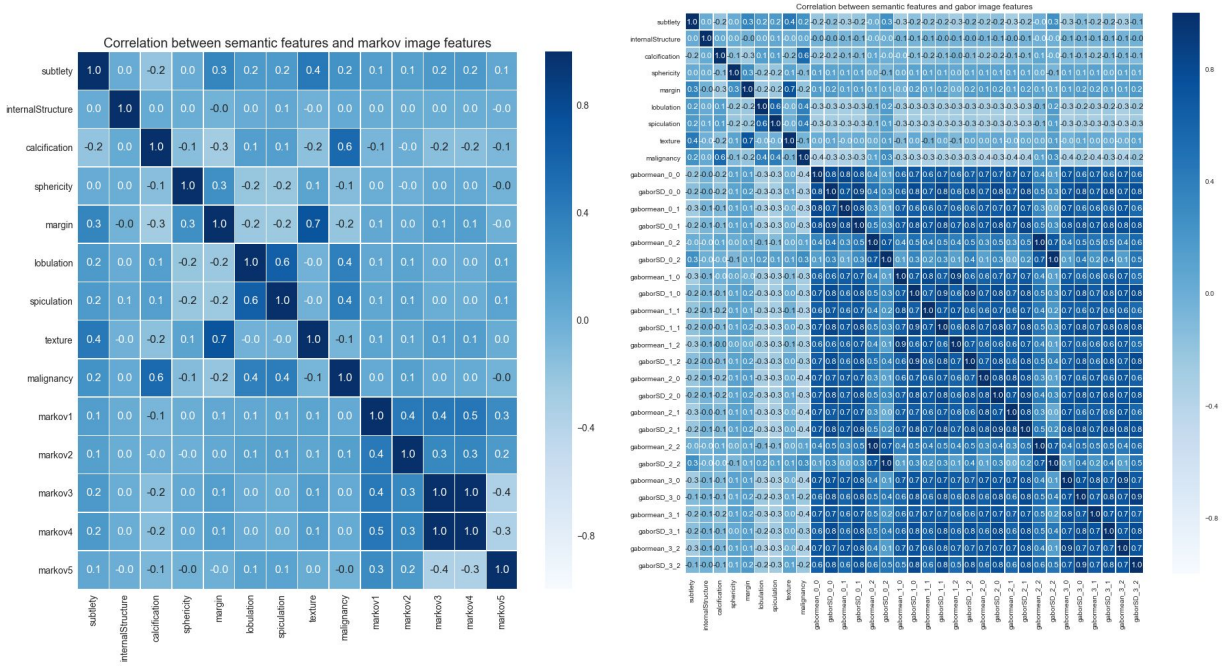
In the correlations between semantic features and different category of image features(fig.11a,b), malignancy shows high correlation(0.6) with image size features, as well as lobulation and spiculation. The spiculation and roughness have 0.5 correlation value. In the correlation of semantic features and intensity image features, calcification has negative correlation with Max Intensity (-0.6), SD Intensity (-0.7), and intensity difference (-0.5). Calcification has highly negative correlation with Contrast (-0.7), Sum Average (-0.5), variance (-0.7), and cluster tendency (-0.7). Entropy has positive correlation with lobulation (0.5), spiculation (0.5) and malignancy (0.6).

Although semantic features and markov image features do not have correlations, markov3 and markov4 have extremely high correlation (1.0). Semantic features and gabor image features also don't display significant correlation.

The image features generally don't display significant correlation with semantic features. However, the image features are correlated amongst themselves, which isn't a surprise as they are calculated with related measurements or have statistical relationships, such as mean and standard deviations.

Among the size image features, the intensity image features and gabor image features, they have high positive correlations (0.9 or 1.0). The shape image features are highly positively and negatively correlated. In the harlick image features, sum average, variance, and cluster tendency have high positive correlation, as well.

Figure 11b. The correlation between semantic features and image features



Because of the high multicollinearity amongst the image features, as well as the high quantity of features, we will use principal component analysis (PCA) to both reduce correlation and number of features.

The PCA plot (fig. 11c) to the right is our first attempt to reduce the number of image features, although a little more fine tuning will be necessary to move forward.

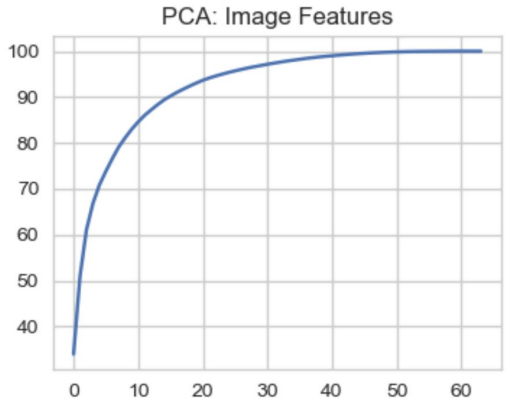


Figure 11c. Plot of Principal Component Analysis of Image Features

Probabilistic Vector Classification

To create the probabilistic malignancy target vector, we will use the following technique:

noduleID	Rad 1	Rad 2	Rad 3	Rad 4	NumRads	NumAgree	Probabilistic Vector
1	1	2	0	0	2	1	[0.5, 0.5, 0, 0, 0]
2	1	1	2	0	3	2	[0.66, 0.33, 0, 0, 0]
3	1	1	1	3	4	3	[0.75, 0, 0.25, 0, 0]
4	1	1	1	1	4	4	[1.0, 0.0, 0, 0, 0]

Other Datasets

In our study, 9 subjective features, rating by 4 independent radiologist experts according to their experiences, gave us the biggest challenge. Medical images, the basic way to get the diagnosis of cancer, must be read by radiologists. There are two common procedure 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 the similar issue. This dataset is cardiovascular MRI of 15 patients. They are read by 7 independent expert readers, according to their own standard operating protocols and their preferred tools. Since the dataset has uncertainty inputs issue, but also multiple readers problem, we consider that our dataset as well as all medical images have the same challenge.

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

References

[1] Yung, M.; Furst, J.; Raicu, D. "Multi-Class Malignancy Prediction with Oversampling Technique Analysis," DePaul University, 2017 (unpublished).

[2] Zinovev D., Raicu DS., Furst JD., and Armatto III, SG. "Predicting Radiological Panel Opinions Using a Panel of Machine Learning Classifiers," *Algorithms*, 2, p1473-1502, 2009.

[3] Zinovev D., Furst JD., and Raicu DS. "Building an Ensemble of Probabilistic Classifiers for Lung Nodule Interpretation", *The tenth International Conference on Machine Learning and Applications* (ICMLA'11), December 18-21, 2011.