**Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features**

**J. Kalpathy-Cramer1, A. Mamomov1, B. Zhao2, L. Lu2, D. Cherezov3, S. Napel4, S. Echegaray4, M. McNitt-Gray5, P. Lo5, J.C. Sieren6, J. Uthoff6,  S.K.N. Dilger6, B. Driscoll7, I. Yeung7, L. Hadjiiski8, K. Cha8, Y. Balagurunathan9, R. Gillies9, D. Goldgof3**

**1Massachusetts General Hospital, 2Columbia University Medical Center , 3University of South Florida, 4Stanford University, 5University of California Los Angeles, 6University of Iowa, 7Princess Margaret Cancer Center, 8University of Michigan, 9Moffitt Cancer Center**

**Abstract:**

Radiomics has been used to provide quantitative descriptors of normal and abnormal tissues during classification and prediction tasks in radiology and oncology. Members of the Quantitative Imaging Network are developing radiomics “feature” sets to characterize tumors. These mathematical descriptors provide ways to characterize the size, shape, texture, intensity, margin and other aspects of the imaging features of nodules and lesions. There is an ongoing effort to develop an ontology to describe radiomics features for lung nodules with the main classes consisting of size, local and global shape descriptors, margin, intensity and texture based features. Texture features were based on wavelets, Laplacians of Gaussians, Law’s features, gray level co-occurrence matrices and run length features.

The purpose of this study was: (a) to investigate the sensitivity of quantitative descriptors of pulmonary nodules to segmentations and (b) to illustrate comparisons across different feature types and features computed by different implementations of feature extraction algorithms. We calculated the concordance correlation coefficients of the features as a measure of their stability to the underlying segmentation. 68% of the 830 features in this study had a concordance correlation coefficient of 0.75 or higher. Pairwise correlation coefficients between pairs of features were used to uncover associations between features, especially as measured by different participants. A graphical model approach was used to enumerate the number of uncorrelated feature groups at given thresholds of correlation. At a threshold of 0.75, there were 75 subgroups while at 0.95, there were 246 subgroups, providing a measure for the redundancy of the features.

1. **Introduction**

Radiomics, “the high-throughput extraction of large amounts of image features from radiographic images” (1), has been used to provide quantitative descriptors of normal and abnormal tissues during classification and prediction tasks in radiology and oncology. Currently, several groups within the Quantitative Imaging Network (QIN) are developing radiomics “feature” sets to characterize tumors. These mathematical descriptors provide ways to characterize the size, shape, texture, intensity, margin and other aspects of the imaging features of nodules and lesions with eventual goals of being able to separate benign from malignant nodules, assessing response to therapy and correlating imaging with genomics. Since tumors usually occupy a relatively small portion of the radiological image or volume, a common requirement for any feature extraction algorithm is dependence on a provided Region of Interest (ROI), commonly referred to as segmentation.

The purpose of this study was twofold: (a) to investigate the sensitivity of quantitative descriptors of pulmonary nodules to segmentations by examining the variability of the features to variations in the segmentations caused by use of different algorithms and/or initial conditions they require, and (b) to illustrate comparisons across different feature types and features computed by different implementations of feature extraction algorithms. This important first step is required to understand feature stability and associations amongst features. However, due to the nature of the data available, in this work, we could not and do not answer questions about the utility of specific quantitative image features for prediction of malignancy, pathological nodule diagnosis, response to therapy or other possible clinically-related questions.

The images used in this study were generated during a previous multi-site study (2) that investigated the repeatability and reproducibility of lung nodule segmentation algorithms in a dataset of 52 lesions in 41 CT volumes, each of which was segmented by three different algorithms. All CT volumes as well as the segmentations are available in the Cancer Imaging Archive (TCIA) as a shared list for easy distribution. Eight sites participated in the current study and seven provided features extracted at each site using these segmentations, as well as “dictionary” files that provided meta-data about the features they computed, while one site provided the infrastructure and statistical analysis support for the project. Ongoing efforts are underway to create an ontology of features based on the information provided in these feature dictionaries.

1. **Methods**

Dataset:

We used 41 CT volumes from 5 collections of DICOM CT images of subjects with Non-Small Cell Lung Cancer and a thoracic phantom previously utilized for the QIN lung segmentation challenge (2) from the following sources: CUMC\_FDA Phantom, Moffitt Cancer Center, the Reference Image Database to Evaluate Therapy Response (RIDER) (3), Stanford University Medical Center, and the Lung Image Database Consortium (LIDC) (4). All image data from human subjects were previously de-identified under IRB protocols in place at participating institutions prior to deployment on TCIA (5) and so, for this purposes of the current study, these data are not considered human subjects data and, hence, no additional IRB supervision is required. The collection consisted of lung CT volumes collectively containing 52 nodules and segmentations. There were 40 cancer patients with a single lesion of interest per scan and 12 phantom nodules all within one additional scan. Nodules varied by location, size and other attributes.  Three different segmentation algorithms were used, with each using three different initial parameters (such as seed point or ROI), to obtain 9 segmentations per nodule, resulting in a total of 468 segmentations.

In the current study, seven groups (University of Columbia, University of South Florida, Stanford University, University of California Los Angeles, University of Iowa, Princess Margaret Cancer Center, University of Michigan) obtained the dataset, including all images and segmentations, from TCIA (5), and each computed their own set of features for each of 468 segmentations and uploaded them to a website set up for this feature challenge. In addition, each participant uploaded a “feature dictionary” that facilitated mapping of each feature to one of a set of predefined feature categories and provided other metadata about the features including a short description.

Feature ontology:

One of the on-going goals for this project is to define an ontology of features that are typically extracted for lung lesions. All participants, through an iterative process, defined a set of feature classes that covered the types of features that were being extracted by this group. Some of the feature classes (e.g., texture) were further subdivided into feature sub-classes. Participants provided information about the feature “class” and “subclass” as part of the dictionary file.

This ontology also facilitated comparisons of features across institutions within a class or subclass. The feature classes agreed upon were: size, intensity, global shape descriptors (GSDs), local shape descriptors (LSDs), margin and texture features. Because of the diversity of texture features, this class was further divided into several subclasses of texture features such as Gray Level Co-occurence Matrix features (also known as Haralick texture features), Laplacian of Gaussian features, Law’s texture features, Run Length features and Wavelet based features. When sites provided their feature dictionary, they described each feature in terms of its name, class, subclass (if applicable) and some general descriptors relating to the calculation of the feature such as whether the calculation was done in two dimensions (2D) or three dimensions (3D), whether it was multiscale or not and, if so, the number of scales. Each participating site provided this dictionary; a brief description of their submitted features including the computational pipeline are provided below.

Feature computation

The extractions for features of lung lesion typically followed a standard sequence of image analysis steps. While nodule segmentations were provided, some algorithms additionally segmented the surrounding parenchyma and/or the lungs. Preprocessing steps varied amongst participants, and included steps such as upsampling the volume to high resolution isotropic spaces and/or image cropping.

Following any preprocessing steps, typical features computed included those related to the size of the lesion (such as volume, maximal diameter, size of bounding box), local (e.g. roughness) and global (e.g. eccentricity) shape descriptors, lesion intensity (e.g., average, median, maximum, minimum, standard deviation voxel values), margin (e.g., edge gradients, surface normals) and texture (e.g., those based on gray level co-occurrence matrices, wavelets).

Some features (e.g., volume, maximal diameter, shape) were calculated based on just the provided nodule segmentation while others were based on the intensity values within the segmentations of the CT volume. Some participants calculated only 3D features, other calculated both 2D (for example, from the maximal trans-axial cross-section of the provided segmentation) and 3D features. Some features (e.g., textures) were calculated at each point within the segmentation and aggregated over the volume. Some of these features were calculated at multiple scales and directions while others were scale and rotation invariant. The aggregation methods varied amongst the participants from simple averaging over scales, directions, locations to more complicated methods such as kernel functions or Gaussian mixtures.

In aggregate, 7 participating sites (**table 1**) provided a total of 830 features. The following is a brief summary of the processing employed by each participating institution along with, in some cases, references to more complete descriptions of the features:

Columbia University Medical Center (CUMC)

The CUMC feature pipeline (6) consisted of two stages, image preprocessing and feature extraction. In the image preprocessing stage, each nodule was firstly cropped out by using a bounding box extending 1 cm beyond the largest nodule extent in all three dimensions, and then linearly interpolated into 3D isotropic images with voxel spacing of 0.5X0.5X0.5 mm. In the feature extraction stage, a set of 71 radiomic features were extracted. Among the 71 radiomic features, some were computed in 2D and some were computed in 3D. The 2D features were calculated on the automatically determined axial image where the nodule had the maximal diameter. In the implementation, when involving parameters of neighborhood, direction and distance between pixels/voxels, 8 connected pixels were considered as the neighboring pixels for 2D analysis, whereas 26 connected voxels were considered as the neighboring voxels for 3D analysis; 8 directions were used for 2D analysis and 13 directions were used for 3D analysis; unless specified, the distance between two neighboring pixels/voxels was one The CUMC feature extraction algorithms were all written in Matlab.

Princess Margaret Cancer Centre (PM)

The PM extracted the 3D lesion by applying the mask to the 3D image set. The extracted image subset was used to calculate the intensity and size sub class features using in-house algorithms developed in Matlab. The only exception was that the 3D images were up sampled to 1 mm slices before being applied to the surface area algorithm which was also written in Matlab.

Stanford University

Stanford employed a prototype of its 3D Quantitative Image Feature Pipeline (QIFP) (7) to compute 3D features from the 468 nodule segmentations.  The software was written using Matlab and takes as input a DICOM Segmentation Object, which unambiguously defines the volume of interest containing the nodule, and a DICOM Image Series consisting of the CT sections acquired by scanning the subject. The QIFP consists of a preprocessing stage, which establishes voxel-to-millimeter scaling factors and crops each nodule with a bounding box extending 2 cm beyond the largest nodule extent in all three dimensions to limit storage and processing requirements. It then computes the following general classes of features: (1) size features: surface area (mm2) and volume (mm3), (2) Intensity features: various statistics of the voxel intensity histogram, (3) General Shape Descriptor: sphericity, (4) Local Shape Descriptors: Roughness statistics and Local Volume Invariant statistics, (5) Margin Features: Statistics of 2-parameter fit to sigmoid function of intensities along surface normals at 800 locations around the nodule, (6) Texture features: mean and standard deviation over orientation of Haralick features (derived from the co-occurrence matrix) at 3 scales (1 mm, 2 mm, 3 mm), for a total of 197 scalar-valued features.  Detailed formulae for these features can be found in (8).

University of California Los Angeles (UCLA)

The UCLA feature pipeline created the reported feature values by first reading in the nodule cases as DICOM images and the nodule regions of interest (ROIs) from segmentations provided. For each nodule and each ROI, fifteen features were calculated and submitted for analysis, which represented a modest subset of features available.  Volume in mm3 was calculated from the number of identified voxels in each ROI and the size of each voxel as reported in the DICOM header. Four additional features were based on the intensity distribution of voxels contained within each ROI including mean, standard deviation, skewness and kurtosis. Finally, 10 features from the Gray Level Co-occurrence matrix (GLCM) texture family were submitted. The co-occurrence matrices were formed using 32 quantization levels (of gray levels or HU bins). The number of directions used (also referred to as offsets) were based on the direction on the 26-connectivity that is typically used in 3D, resulting in a total of 13 offsets (= 26/2 due to symmetry). This results in a total of 13 co-occurrence matrices. The measures calculated (contrast, dissimilarity, homogeneity, energy and entropy were used for this study) were obtained for each co-occurrence matrix formed and the final feature values submitted were the mean value and range value (maximum – minimum) of each measure over the 13 co-occurrence matrices.

University of Iowa (UIowa)

The UIowa feature pipeline consisted of image preprocessing and feature extraction, which were both done using scripts written in Matlab. The UIowa approach included feature extraction from both the lung nodule and the surrounding lung parenchyma. For this study the nodule segmentation mask was provided, from which the maximum radius of the nodule was determined. Extending from the nodule boundary, all voxels within a distance of the maximum nodule radius were included in the parenchymal mask. If non-valid parenchyma regions were present, such as pleural wall, they were manually excluded from the parenchyma mask. In the feature extraction stage, a set of 304 radiomic features were calculated as previously described in (9). From the nodule, 159 features were extracted including intensity, size, shape, and texture.  From the surrounding parenchyma mask, 145 features were extracted including intensity and texture features, which are reported here as margin features. 2D and 3D shape and size features were calculated using physical distance, to compensate for voxel resolution differences in the dataset. Texture features were calculated based on a three-dimensional, rotation invariant implementation of Law’s texture energy measures (10) aggregated as histogram summary statistics.

University of Michigan (UMICH)

The general Michigan feature extraction pipeline consists of 3 stages: (1) preprocessing, (2) nodule segmentation, (3) feature extraction. It was developed using C and it takes DICOM images as input.  In the preprocessing stage, a lung mask is first generated by segmenting the lungs using k-means clustering.  By using an input bounding box around the target nodule (with a margin of about 10 mm around the nodule), a volume of interest (VOI) containing the nodule is extracted. An image of isotropic resolution is obtained by performing linear interpolation in x, y, z direction. The nodule segmentation stage (Stage 2) was not activated in this specific challenge because the nodule segmentations were provided and they were imported into the pipeline. The imported nodule segmentations were positioned on the interpolated slices by selecting the nodule segmentation nearest to each interpolated slice. In the feature extraction stage (Stage 3), the size, intensity, texture, shape, and margin features were extracted. The mediastinal or pleural voxels were excluded from the feature extraction process using the lung mask. Within the segmented regions, size features such as volume, surface area, perimeter, and diameter, as well as intensity features, including average, variance, skewness, and kurtosis of the gray-level histogram were extracted (11). After performing the rubber band straightening transform (RBST) (12) of the segmented nodules slice by slice, texture features based on run-length statistics (run length features) are extracted from the Sobel-filtered RBST images (11). In addition, margin features based on the statistics of gradient field strength and orientation on spherical shells around the nodule surface were extracted to describe the sharpness of the nodule boundary and the smoothness of the nodule margin (13, 14).  Finally, global shape descriptors based on the statistics of nodule radii were extracted as descriptors of the irregularity of the nodule shape (13).

Moffitt Cancer Center/University of South Florida (USF):

The Moffitt Cancer Center/ University of South Florida teams workflow was implemented on a custom implementations of imaging features in C/C++ language that could be called from a commercial imaging workstation (15). Quantitative imaging features were extracted on these regions of interest. The USF pipeline has only one stage. There were no preprocessing operations applied on the DICOM images or the segmentations. The in-plane pixel spacing parameters are typically defined during the scan, which is dependent on the patient size and or the scan center. The USF methods had no modifications applied to segmentation or DICOM files. The implementation had 184 features in total, categorized into size, density intensity, shape, margin and texture (co-occurrence, run length, wavelets & Law’s features). All the features were computed in 3D. These features had been qualified for their reproducibility on a test/retest dataset (16). These features have been shown to be effective to predicting malignancy in screening setting (17) and indicator of disease progression and related to the genome (18, 19).

Statistical analysis

Our first goal was to understand the sensitivity of the features to the segmentation. We used the repeated measures Concordance Correlation Coefficient (CCC) (20) as our statistical estimate of the repeatability and reproducibility of our features to the segmentations. We calculated inter-segmentation algorithm, intra-segmentation algorithm and total (combining intra- and inter-) CCC for each feature.

Our second goal was to understand the correlations amongst features collected by the 7 participants. For this, we calculated the association using the correlation coefficients (CC) between all pairs of features. We expect similarly named features (e.g. volume) to be highly correlated across different participant’s implementations. However, we were also interested looking at correlations within feature sets provided by each participating institution, and in identifying unique, uncorrelated features both within and across all participating institutions.

A graphical model approach was utilized to examine the correlations among features. Each of the 830 features was modeled as node in an undirected graph and the edge weight between two nodes was the absolute value of the correlation coefficient. We used both the Pearson and Spearman correlation coefficients, as associations between features can be linear or non-linear. Starting with a fully connected graph, edges were then filtered such that edges with weight less than given threshold (T) were removed. This produced sets of disjoint sub-graphs. We computed the number of such sub-graphs (at least one node) for a number of different thresholds as a measure of the uniqueness of the features submitted.

1. **Results:**

Feature Ontology: The number of features from each participant varied from 10 to 304 and **table 1** enumerates the features in each class and sub-class per the ontology used for this project, based on the meta-data provided with each submission.

Table 1. Number of features per type submitted by each participating institution.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Participant** | **Size** | **GSDs** | **LSDs** | **Intensity** | **Margin** | **Texture** | | | | | **Total** |
|  |  |  |  |  |  | **GLCM** | **LoG** | **Law's** | **Run length** | **Wavelet** |  |
| CUMC | 3 | 4 | 8 | 5 |  | 17 | 6 | 14 |  | 14 | 71 |
| PM | 3 | 2 |  | 5 |  |  |  |  |  |  | 10 |
| Stanford | 2 | 1 | 78 | 17 | 27 | 72 |  |  |  |  | 197 |
| UCLA | 1 |  |  | 4 |  | 10 |  |  |  |  | 15 |
| UIowa | 2 | 6 |  | 9 | 151 |  |  | 136 |  |  | 304 |
| UMICH | 4 | 5 |  | 6 | 18 |  |  |  | 16 |  | 49 |
| USF | 5 |  |  |  |  | 6 |  | 125 | 20 | 28 | 184 |
| Total | 20 | 18 | 86 | 46 | 196 | 104 | 6 | 275 | 36 | 42 | 830 |

Data visualization: We have a developed a web-based interactive results viewer [link to be provided] as part of the platform used for this challenge. This cross browser tool, developed using JavaScript allows users to explore a range of analyses of this data. Visualizations include histograms of features by site and feature class, heat maps of correlations of features between sites, and graphical models of the connectivity of features.

Statistical analysis: We calculated the inter-, intra- and overall CCC for each feature. High CCC imply that the features are not very sensitive to the underlying segmentation while low CCCs suggest that the characteristics of the underlying segmentation have a strong influence on the value of that features. The inter CCC is an estimate of the stability of the feature across different segmentation algorithms while the intra CCC is the estimate of the stability of the feature within multiple segmentations of the same segmentation algorithm (with different initializations). In our previous work (2), we demonstrated that the segmentations had higher repeatability than reproducibility (i.e. were more similar among different runs of the same algorithm compared to runs of different algorithms).

**Figure 1** displays the range of overall CCC for the features and highlights the relative stability of a majority of features to variations in segmentation algorithms. ANOVA analysis of the inter, intra and overall CCC indicated that the CCC is different by feature class (p=0.000125 for overall CCC, p<<0.05 for inter CC and p=0.0395 for intra CCC). Further, post-hoc comparisons based on the Tukey-HSD statistic suggested that the inter and overall GSD CCCs are lower and different from the other classes at an adjusted p-value of 0.05 for the inter and overall comparisons.

**Figure 2** displays cumulative density function of the overall CCC by feature class. 68% of all features have an overall CCC of 0.75 or higher. However, just 50% of GSD features have a CCC of 0.75 or higher while 95% of size features have a CCC of 0.75 or higher.

It is worth mentioning that although the boxplot in **figure 1** treats each feature as an independent measure, many features are expected to be highly correlated or identical. For example, we would expect identical features (e.g. volume) that are part of the feature pipelines of multiple participants to be highly correlated. Features within the same class (e.g. GLCM features at different directions or scales) are also potentially correlated.

We also calculated the correlation coefficient for all pairs of features and validated the assumption that the same feature calculated by multiple participants would exhibit high correlation. The most obvious example of such a feature is the tumor volume, a member of the size class; as computed by six of the participants the correlation coefficients between all pairs were between 0.9999 and 1. Other common features such as intensity based mean, standard deviation, median, kurtosis and skewness were calculated by many participants, and were highly correlated between many pairs of participants. Features from a similar class, such as texture features based on GLCMs, were also highly correlated amongst themselves.

**Table 2** displays the number of non-correlated subgroups as a function of the threshold of the CC used for connectedness. At a given threshold, each sub-group corresponds to correlated features and the number of resulting sub-graphs captures approximate dimensionality of overall features. We examined the composition of some exemplar connected subgraphs at different edge strengths. As expected, a lower threshold results in fewer subgroups while a high threshold resulted in a large number of groups. Groups can be seen as a function of both participating institution and feature class in our interactive website, developed as part of our challenge platform. **Figure 3** is an example of the graph at a threshold of 0.95, where the nodes are colored by feature class (above) or site (below). Only groups with more than one node are shown for clarity. Some sub-groups consist of nodes from a single participant while others have representation from many groups. Example groups that have representation from multiple participants include volume, radius, histogram mean, maximum, kurtosis and skewness. **Figure 4** displays the graph at a threshold of 0.75. As expected, the number of nodes per group is larger with more complex structure.

Six of the features in the skewness sub-group were related to the histogram skewness. However, one feature was the entropy mean, which in turn was connected to a run length (subclass), texture (class) feature. This graphical tool allowed us to explore both expected and unexpected clusters (e.g. features from different classes) of features at different levels of connectedness.

Table 2. Number of non-correlated subgroups by given threshold and correlation type. T is the correlation threshold.

|  |  |  |  |
| --- | --- | --- | --- |
| Linear | | Non-Linear | |
| T | # of subgroups | T | # of subgroups |
| .75 | 75 | .75 | 58 |
| .80 | 103 | .80 | 80 |
| .85 | 150 | .85 | 120 |
| .90 | 245 | .90 | 172 |
| .95 | 382 | .95 | 246 |

1. **Discussion**

This study was primarily designed to examine the sensitivity of quantitative image features to tumor segmentation. This was done by examining individual feature values produced by software instances from 7 independent participating institutions across 9 different image segmentation results for each of the 52 lung nodules in chest CT scans. This study does not address the ultimate utility of these feature values in predicting or assessing outcome measures such as assessing whether a lung lesion is benign or malignant or whether the patient is responding to therapy or not. However, unless segmentation algorithms perform perfectly and reproducibly, it is important to understand the stability of any feature to segmentation that could be considered for prediction of these types of clinical variables.

However, in all of the above contexts, region of interest identification (segmentation) becomes most critical. Stability of segmentation algorithms to perform reproducibly well becomes important. Our prior work has shown acceptable repeatability and reproducibility of segmentations across institutions (2). In continuation, understanding the variability of any feature derived on these segmentation becomes ever more critical as the metrics are typically used to relate to the clinical variables and or track response to treatment.

Because of the variety of segmentation algorithms available and because of the variability of any non-fully automated algorithms, sensitivity to segmentation algorithm results is an important issue. However, it is recognized that this is not the only issue that needs to be addressed for understanding the ultimate utility radiomics features; even features that can be shown to be stable over a wide range of segmentation results may not necessarily be useful in performing a given outcome-related task such as differentiating benign nodules from malignant ones or in assessing response to therapy. For example, a feature that yields that same value across all segmentations may be insensitive to changes in biological status and therefore not helpful in assessing treatment response at all. Conversely, features that are sensitive to segmentation results do not necessarily mean that they are not useful in an outcome related task. For example, if the size of the lesion is to be used in assessing treatment response, then having a size feature that does vary with segmentation results would be desirable. Thus, assessing the sensitivity of features across segmentation results is an important issue, but it is not the only issue to be considered when features are being computed for some outcome-related task. In a future study, we plan to investigate the ability of these features to differentiate between benign and malignant nodules in a lung screening CT context.

While the results shown here did show high correlations between certain groups of features as expected (e.g. size features calculated across participants), there were also specific features within these groups that did not. One example was within the intensity feature category; while most participants reported the mean intensity (HU) of each lesion, some participants additionally reported unique and uncorrelated features, such as the size of airspaces within the lesion or the lesion’s maximum or minimum intensity values. While these features may be intensity related, they may not be highly correlated to the mean intensity value of the total nodule and therefore it is not unexpected that these features would produce values that would be shown as outliers in a distribution of intensity features. That said, these features may contribute information that is complementary to the information provided by the other intensity features and may be able to contribute to the outcome-related task (e.g., discrimination or assessment of response).

There were several lessons learned from this study. First of all, it demonstrated that there is substantial value in comparing feature values among different groups, even when the feature values are expected to be the same or very similar. For example, by comparing lesion volume values across different lesions, segmentations and participants, we were able to uncover subtle differences and even errors in approach and calculations that may not have been discovered otherwise; including how participants were handling cases where the slice thickness and slice spacing were not the same value (e.g., overlapping images).

This also illustrated the value of using phantom images or synthetic images where there are objects with known values such as known density or known volume. These allow users the ability to gain confidence that their methods and calculations are performing in a manner similar to some reference method(s). These are helpful steps that can and should be taken when possible before moving on to more complex objects such as the lung lesions used in this study. In this study, having a common set of reference images, well specified objects and existing object masks, allowed the authors to focus on the very specific task of feature computation, its sensitivity to segmentation results, and the associations among specific features.

**Acknowledgments:**

JKC and AM were supported in part through funding from NIH/NCI (U24CA180927, U24CA180918, U01CA154601).

BZ and LL were supported in part through funding from NIH/NCI (U24CA180927, R01 CA149490 and U01 CA140207)

MMG and PL were supported in part through funding from NIH/NCI (U01 CA181156).

SN and SE were supported in part through funding from NIH/NCI (R01 CA160251 and U01 CA187947).

JCS and SKND were supported in part through funding from NIH/NHLBI (R01HL112986).

LH and KC were supported in part by NIH/NCI (U01CA179106). Large part of the general Michigan feature extraction pipeline was developed under the support of NIH award number R01 CA93517 (PI: Heang-Ping Chan). We would also like to acknowledge Dr. Heang-Ping Chan for her substantial contribution to the Michigan team in the QIN Feature challenge.

The work of PMCC was partially supported by the Canadian Institute of Health Research.

DC, DG, YB and RG were supported in part through funding from NIH/NCI (1U01 CA 143062-01) and State of Florida James & Esther King biomedical Research Program, (2KT01).

**References**

1. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, Zegers CML, Gillies R, Boellard R, Dekker A, Aerts HJWL. Radiomics: Extracting more information from medical images using advanced feature analysis2012;48(4):441-6. doi: 10.1016/j.ejca.2011.11.036. PubMed PMID: 22257792.

2. Kalpathy-Cramer J, Zhao B, Goldgof D, Gu Y, Wang X, Yang H, Tan Y, Gillies R, Napel S. A Comparison of Lung Nodule Segmentation Algorithms: Methods and Results from a Multi-institutional Study. J Digit Imaging. 2016;29(4):476-87. doi: 10.1007/s10278-016-9859-z. PubMed PMID: 26847203; PMCID: 4942386.

3. Armato SG, 3rd, Meyer CR, McNitt-Gray MF, McLennan G, Reeves AP, Croft BY, Clarke LP, Group RR. The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: a resource for the development of change-analysis software. Clinical pharmacology and therapeutics. 2008;84(4):448-56. doi: 10.1038/clpt.2008.161. PubMed PMID: 18754000; PMCID: 4938843.

4. Armato SG, 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, Van Beeke EJ, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais RC, Qing DP, Roberts RY, Smith AR, Starkey A, Batrah P, Caligiuri P, Farooqi A, Gladish GW, Jude CM, Munden RF, Petkovska I, Quint LE, Schwartz LH, Sundaram B, Dodd LE, Fenimore C, Gur D, Petrick N, Freymann J, Kirby J, Hughes B, Casteele AV, Gupte S, Sallamm M, Heath MD, Kuhn MH, Dharaiya E, Burns R, Fryd DS, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft BY. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics. 2011;38(2):915-31. doi: 10.1118/1.3528204. PubMed PMID: 21452728; PMCID: 3041807.

5. Kalpathy-Cramer J, Napel S, Goldgof D, Zhao B. QIN multi-site collection of Lung CT data with Nodule Segmentations [cited 2015].

6. Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, Schwartz LH. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep. 2016;6:23428. doi: 10.1038/srep23428. PubMed PMID: 27009765; PMCID: 4806325.

7. Napel S, Echegaray S, Gude D, Gevaert O, Rubin DL, editors. The Quantitative Image Feature Pipeline (QIFP) for Discovery, Validation, and Translation of Cancer Imaging Biomarkers. Radiological Society of North America Annual Meeting, ; 2016 December 2016; Chicago, IL.

8. Echegary S. Techical Report Stanford University School of Engineering. 2016.

9. Dilger SK, Uthoff J, Judisch A, Hammond E, Mott SL, Smith BJ, Newell JD, Jr., Hoffman EA, Sieren JC. Improved pulmonary nodule classification utilizing quantitative lung parenchyma features. Journal of medical imaging. 2015;2(4):041004. doi: 10.1117/1.JMI.2.4.041004. PubMed PMID: 26870744; PMCID: 4748146.

10. Laws KI. Rapid texture identification In: Wiener TF, editor. Image Processing for Missile Guidance: SPIE; 1980. p. 376-80.

11. Way TW, Hadjiiski LM, Sahiner B, Chan HP, Cascade PN, Kazerooni EA, Bogot N, Zhou C. Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. Medical physics. 2006;33(7):2323-37. doi: 10.1118/1.2207129. PubMed PMID: 16898434; PMCID: 2728558.

12. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. Medical physics. 1998;25(4):516-26. doi: 10.1118/1.598228. PubMed PMID: 9571620.

13. Way TW, Sahiner B, Chan HP, Hadjiiski L, Cascade PN, Chughtai A, Bogot N, Kazerooni E. Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. Medical physics. 2009;36(7):3086-98. doi: 10.1118/1.3140589. PubMed PMID: 19673208; PMCID: 2832039.

14. Ge Z, Sahiner B, Chan HP, Hadjiiski LM, Cascade PN, Bogot N, Kazerooni EA, Wei J, Zhou C. Computer-aided detection of lung nodules: false positive reduction using a 3D gradient field method and 3D ellipsoid fitting. Medical physics. 2005;32(8):2443-54. doi: 10.1118/1.1944667. PubMed PMID: 16193773; PMCID: 2800987.

15. toolbox L. Definiens Incorporated. Available from: <http://www.defiiniens.com>.

16. Balagurunathan Y, Gu Y, Wang H, Kumar V, Grove O, Hawkins S, Kim J, Goldgof DB, Hall LO, Gatenby RA, Gillies RJ. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Translational oncology. 2014;7(1):72-87. Epub 2014/04/29. PubMed PMID: 24772210; PMCID: PMC3998690.

17. Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, Li Q, Cherezov D, Gatenby RA, Balagurunathan Y, Goldgof D, Schabath MB, Hall L, Gillies RJ. Predicting malignant nodules from screening CTs. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer. 2016. Epub 2016/07/17. doi: 10.1016/j.jtho.2016.07.002. PubMed PMID: 27422797.

18. Grove O, Berglund AE, Schabath MB, Aerts HJ, Dekker A, Wang H, Velazquez ER, Lambin P, Gu Y, Balagurunathan Y, Eikman E, Gatenby RA, Eschrich S, Gillies RJ. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. PloS one. 2015;10(3):e0118261. Epub 2015/03/05. doi: 10.1371/journal.pone.0118261. PubMed PMID: 25739030; PMCID: PMC4349806.

19. Liu Y, Kim J, Balagurunathan Y, Li Q, Garcia AL, Stringfield O, Ye Z, Gillies RJ. Radiomic Features Are Associated With EGFR Mutation Status in Lung Adenocarcinomas. Clinical lung cancer. 2016. Epub 2016/03/28. doi: 10.1016/j.cllc.2016.02.001. PubMed PMID: 27017476.

20. King TS, Chinchilli VM, Carrasco JL. A repeated measures concordance correlation coefficient. Stat Med. 2007;26(16):3095-113. doi: 10.1002/sim.2778. PubMed PMID: 17216594.

**Figure Legends**

Figure 1 Overall CCC by feature class indicates the relative robustness of features to underlying segmentation

Figure 2 Cumulative histogram of overall CCC by feature class\

Figure 3 Example graphs of connectivity between feature nodes using a threshold of 0.95 for the CC highlight correlation between features from different participants

Figure 4 Graphical model of connected components at a CC of 0.75

**Figures**

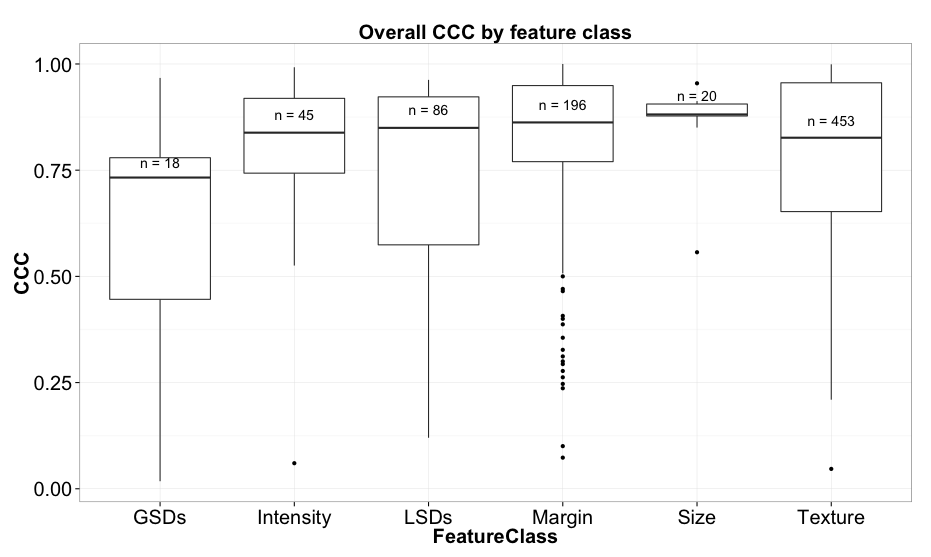


Figure 1 Overall CCC by feature class indicates the relative robustness of features to underlying segmentation

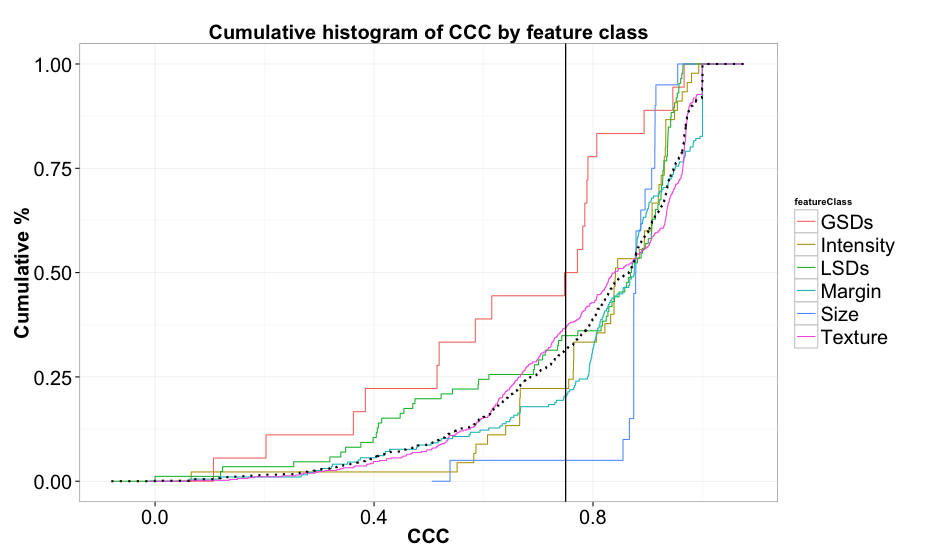


Figure 2 Cumulative histogram of overall CCC by feature class

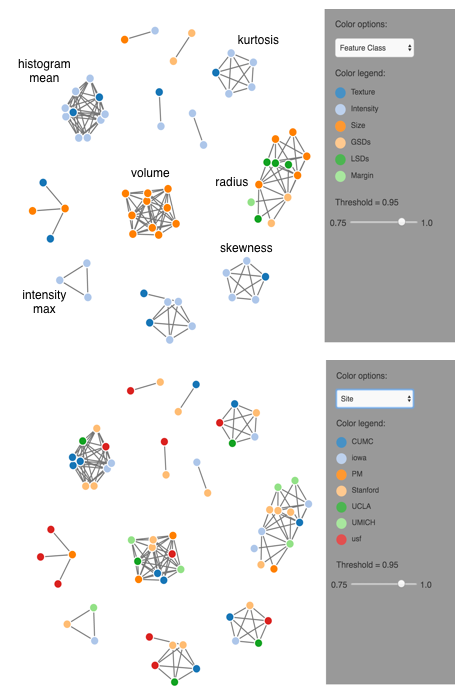


Figure 3 Example graphs of connectivity between feature nodes using a threshold of 0.95 for the CC highlight correlation between features from different participants

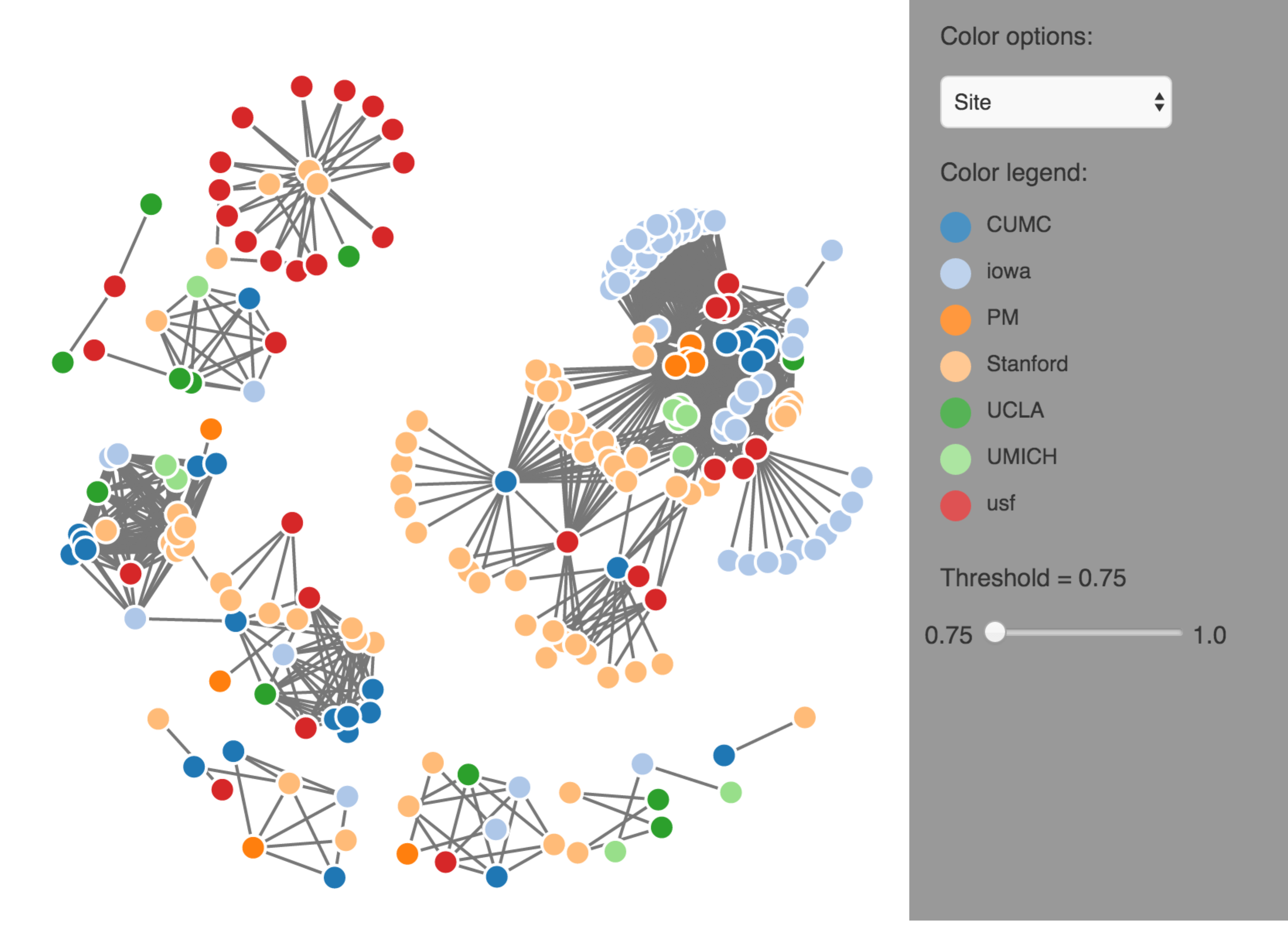
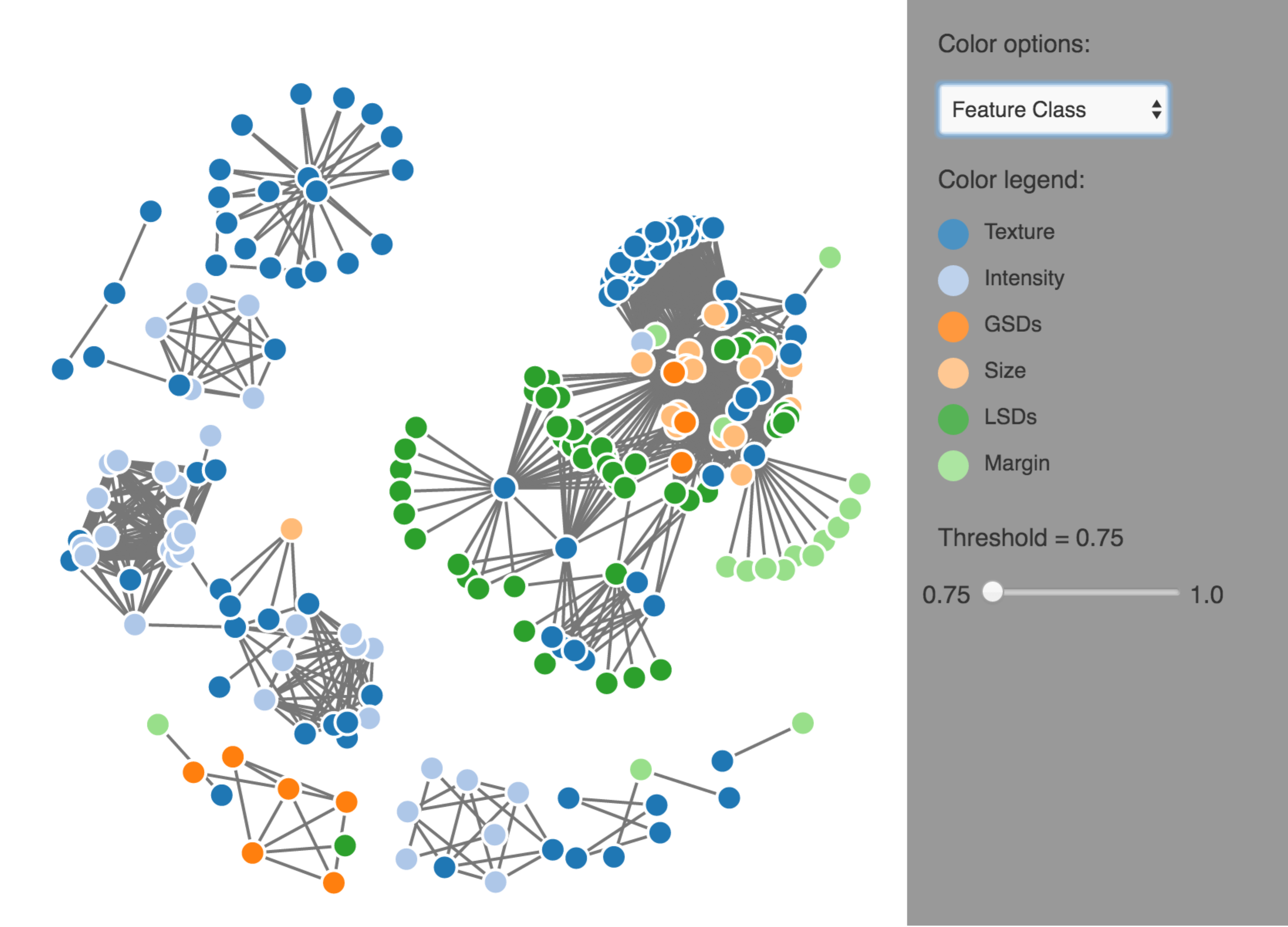


Figure 4 Graphical model of connected components at a CC of 0.75