SESUG 2022 Paper 222

Application of Feature Selection and Dimension Reduction Techniques on Large-Scale CT Dataset for Lung Cancer Diagnosis Based on Radiomics

Mostafa Zahed, East Tennessee State University; Maryam Skafyan, University of Northern Colorado

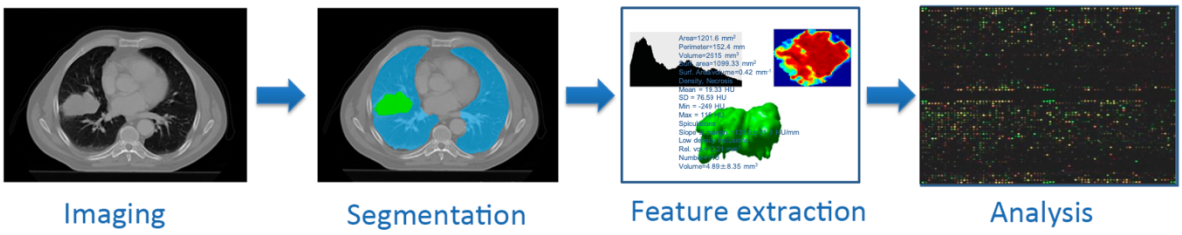
ABSTRACT

A German physicist Wilhelm Konrad Rontgen started to work on medical imaging in 1895. Later, in 1978, Hounsfield developed computed tomography (CT) technology to study medical images to screen and detect tumors in the fastest way. Because of the importance of searching medical images, Lambin 2012 proposed an approach to analyzing medical images, which is called radiomics. This approach has involved several steps from the beginning to the end: segmentation, feature extraction, feature selection, and statistical modeling. Extracted features can be categorized in the description of tumor gray histograms, shape, texture features, and the tumor location and surrounding tissue. Because of the massive number of features from radiological images, machine learning plays an important role in analyzing the big data obtained from tumors. In other words, the dimension of extracted features needs to be reduced to describe the tumor better. Many linear and nonlinear dimension reduction techniques, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDS), Local Feature Analysis (LFA), and manifold learning have been developed. In this paper, a large-scale CT dataset for Lung cancer diagnosis (Lung- PET-CT-Dx) which was collected by Huiping Han, Funing Yang, and Rui Wang of Harbin from Medical University in Harbin in China is used to illustrate the dimension reduction techniques, which is a main part of radiomics process, via SAS. This dataset consists of CT DICOM images of lung cancer subjects with XML Annotation files that indicate tumor location with bounding boxes. Pyradiomics through 3D Slicer medical software was used to extract features for 74 patients out of 130 as the provided annotation file did not work for all patients through Python. This study has been shown how to apply the dimension reduction methods available in SAS to the lung cancer dataset, including Principal Component Analysis (PCA) and Cluster Analysis through PROC FACTOR, PROC PRINCOMP, and TEXTMINE Procedure. Both approaches suggested if the data were categorized into six subcategories, the reduced data would adhere to the dominant variation in the original data.

Keywords: Radiomics, Segmentation, Dimension Reduction, Features Extraction

INTRODUCTION

Radiomics is a method that extracts a large number of features from radiographic medical images using data-characterization algorithms. Radiomics hypothesizes that the distinctive imaging features between disease forms may be useful for predicting prognosis and therapeutic response for various conditions, thus providing valuable personalized therapy information is main goal behind radiomics (Mohammadi et al., 2019). The use of medical imaging technologies is currently entering into the context of individualized medicine (Lambin et al., 2012). There are different sources of information, e.g., demographics, pathology, toxicity, biomarkers, genomics, and proteomics that can be used for selecting the optimal treatment, so it is expected that the image contains more information in comparison of other sources (Lambin, et al., 2010). In the workflow of radiomics, on the medical images, segmentation is performed to define the tumor region. From this region the features are extracted based on tumor intensity, texture, and shape. Finally, these features are used for analysis, Figure 1. The most widely used imaging modality in radiomics is Computed tomography (CT).



*Lambin *et al.* Eur J Cancer 2012
*Aerts *et al*, Nature Comm 2014

Figure 1. The Radiomics workflow

The CT accesses tissue density, shape and texture of tumor (Aerts et al., 2014). CT has also the potential to measure size and change in size during and after therapy in order to early access treatment responses. Positron Emission Tomography (PET) is also used for detecting and staging cancer, most commonly with the radiotracer 18F-fluorodeoxyglucose (18F-FDG), a radiolabeled sugar (glucose analog) molecule (Ypsilantis et al., 2015). PET has been a hotbed for radiomics application to the functional nature of the images and their direct relationship to underlying tumor biology (Naqa, 2014). The calculation of radiomic features is typically applied to a specific (Region of Interest) ROI which needs to be segmented from the surrounding tissues, such as the gross tumor. In most radiomics studies, the tumor is manually delineated by an experienced radiologist or radiation oncologist (Wang et al., 2016). The goal of radiomics is to develop a function or mathematical model to classify patients according to their predicted outcome by means of radiomic features. In the language of pattern recognition machine-learning, this task is equivalent to building a "classifier", which is an algorithm analyzing training data and inferring a hypothesis (the function), to predict the labels of unseen observations, e.g., patient outcome or tumor phenotype (Parmar et al., 2015 and Bannach et al., 2017).

In general, there are several nonlinear and linear approaches for dimension reduction techniques that have been developed. Among them, we can mention Principal component analysis (PCA), Clustering, Local fisher's discriminate analysis (LFDA), Canonical correlation analysis (CCA), Non-negative matrix factorization (NMF), and Manifold learning-based

algorithm (Ray et al. 2021) contrastive Principal Component Analysis (cPCA), JIVE, Group component analysis for multi block data, Joint and Individual Variation Explained (JIVE) for Block Data, Common Orthogonal Basis Extraction (COBE). The goal behind these techniques is to investigate how we can tackle or extend these approaches in analyzing Images, Tissue Slide Images, Clinical Data, Biomedical Data, and Genomics simultaneously.

Contrastive principal component analysis is a method which identifies low-dimensional structures that are enriched in a dataset relative to comparison data (Abid et al., 2018). The main advantages of cPCA are its generality and ease of use. Computing a particular cPCA takes essentially the same amount of time as computing a regular PCA. The Joint and Individual Variation Explained (JIVE) is a method in aim of variation decomposition for the integrated analysis of cancer data. The decomposition consists of three terms: a low-rank approximation capturing joint variation across data types, low-rank approximations for structured variation individual to each data type, and residual noise (Lock et al., 2013). Furthermore, the JIVE has been extended to apply for multi-block data in order to capture both joint and individual variation within each data block (Feng et al., 2018). An R package, R.JIVE, is introduced to perform JIVE and visualize the results (O'Connell and Lock, 2016). We figured out that one should load caTools first for installation of this package. This part did not mention in the installation process provided by (O'Connell and Lock, 2016). Another approach is an efficient algorithm called Common Orthogonal Basis Extraction (COBE) to extract the common basis, which is shared by multi-block data, independent on whether the number of common components is given or not (Zhou et at., 2015).

In SAS settings, it is challenging to apply some of machine learning techniques as this part is still under processing and developing. Consequently, we are focus on very common dimension reduction techniques as such PCA and Clustering. Both tools are a technique for data pre-processing before applying supervised techniques.

In the following, software tools for extracting features are described to explain how many options are available to extract features from tumor in cancer medical images. Then, we describe how the extracted features are organized to prepare the matrix data to be analyzed through unsupervised techniques mentioned above.

SOFTWARE TOOLS FOR EXTRACTING FEATURES

The general Radiomics workflow includes image acquisition, segmentation, features extraction of high-dimensional datasets. There is various software for extracting the features, including PyRadiomcs, 3D Slicer, LIFEx, IBEX, QIFE, and RayPlus. Each tool is attached with its limitations and restrictions, and there is no specific reason to determine which one is the best. Indeed, it depends on how a researcher wants to tackle medical images. After investigation, we found LIFEx and 3D Slicer are more suitable for our needs. Concretely, there are a few good reasons why LIFEx and 3D Slicer should use for features extraction.

PyRadiomics was developed by Griethuysen et al., 2017, to address the lack of standardized algorithm definitions in image processing. PyRadiomics is a flexible open-source platform capable of extracting a large panel of engineered features from medical images. PyRadiomics is implemented in Python and can be used standalone or using 3D Slicer.

Although, the main advantage of this tool is using *SimpleITK*, which makes it very helpful for a variety of image formats. However, the processing of PyRadiomics should be done through Python. It means to use this platform; one needs to have enough knowledge in Python and programming.

3D Slicer is a free, open-source software application for medical image computing. As a clinical research tool, 3D Slicer is similar to a radiology workstation that supports versatile visualizations and provides advanced functionality such as automated segmentation and registration for a variety of application domains. Unlike a typical radiology workstation, 3D Slicer is free and is not tied to specific hardware. As a programming platform, 3D Slicer facilitates translation and evaluation of the new quantitative methods by allowing the biomedical researcher to focus on implementing the algorithm and providing abstractions for the common tasks of data communication, visualization, and user interface development. Compared to other tools that provide aspects of this functionality, 3D Slicer is fully open source and can be readily extended and redistributed. Besides, 3D Slicer is designed to facilitate the development of new functionality in 3D Slicer extensions (Fedorov et al., 2012). But this software is not intended for clinical use; it just works in a panel of medical image informatics, image processing, and three-dimensional visualization, Figure 2.

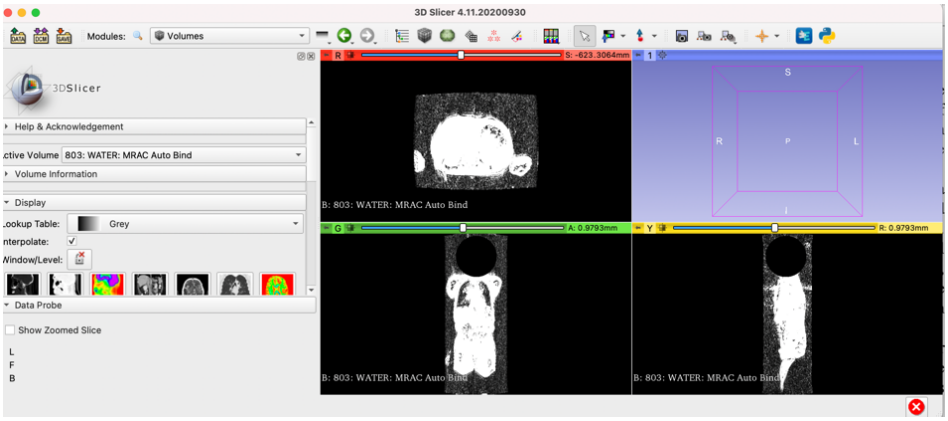


Figure 2. Loading Lung-CT-PET Images, available on The Cancer Imaging Archive (TCIA) Public Access dataset, by 3D Slicer.

DATA PREPARATION FOR UNSUPERSVISED TECHNIQUES

The data is used in this study was collected by Huiping Han, Funing Yang, and Rui Wang of Harbin from Medical University in Harbin in China (Wang et al., 2020). This dataset consists of CT and PET-CT DICOM images of 130 patients with lung cancer. The XML annotation files which includes the location of tumor was provided by five academic radiologists with high expertise in lung cancer. To visualize the annotation boxes on tumor of the DICOM images the following Python through terminal was used to pull out the images and put the box on top of them, Figure 3.

```
(base) mostafa@UM-C02F3165Q05N ~ % cd
/Users/mostafa/Documents/Mostafa/UM/Papers/Data/TCIA/Lung-PET-CT-
Dx/VisualizationTools
```

```
(base) mostafa@UM-C02F3165Q05N VisualizationTools % conda create -n dcm-vis
python=3.7
```

```
(dcm-vis) mostafa@UM-C02F3165Q05N VisualizationTools % python
visualization.py --dicom-mode="CT" --dicom-
path="/Users/mostafa/Documents/Mostafa/UM/Papers/Data/TCIA/Lung-PET-CT-
Dx/VisualizationTools/Data/Images/manifest-1608669183333/Lung-PET-CT-
Dx/Lung_Dx-A0239" --annotation-
path="/Users/mostafa/Documents/Mostafa/UM/Papers/Data/TCIA/Lung-PET-CT-
Dx/VisualizationTools/Data/Annotation/A0239" --
classfile="/Users/mostafa/Documents/Mostafa/UM/Papers/Data/TCIA/Lung-PET-CT-
Dx/VisualizationTools/category.txt"
```

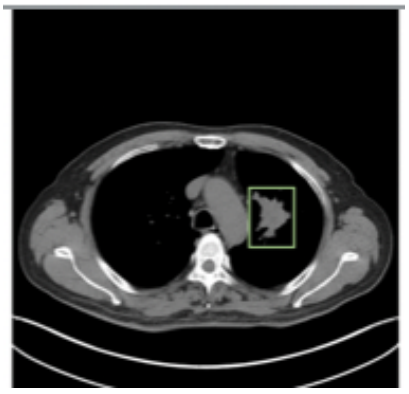


Figure 3. Visualization of the annotation box on the CT-DICOM images.

After implementing the provided annotation files, we realized that it just worked for 74 patients out 130. So, the number of rows for the target matrix data was 74. After this stage, 3D Slicer was used to do the segmentation process and by using PyRadomics package available in 3D Slicer, the features were extracted for each patient. The dimension of the obtained matrix data was 74 by 110. Each row represented the patient, and each column was for the extracted feature. The extracted feature was categorized in three types: 3D features, Texture features, and Intensity features. For instance, the texture features are mainly used to explain the nature of the pixels of Region of Interest (ROI) and its surrounding pixels. It can be computed by using the Gray-Level-Co-occurrence Matrix (GLCM), Gray-Level Size Zone Matrix (GLSZM), Gray-Level Run Length Matrix (GLRLM), Neighborhood Gray-Ton Difference Matrix (NGTDM), and Gray-Level Dependence Matrix (GLDM), Figure 4.

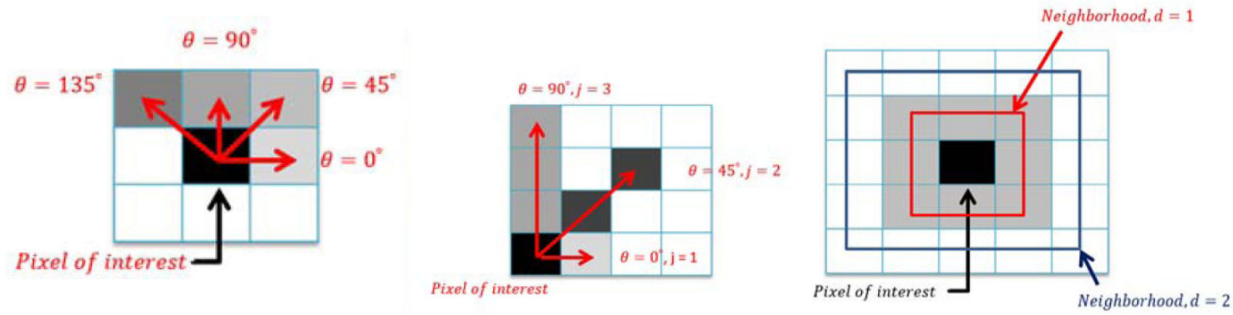


Figure 4. Texture Features from left to right: GLCM (gray level co-occurrence matrix), GLRLM (gray level run length matrix) and NGTDM (neighborhood gray tone difference matrix) (Parekh and Jacobs (2016))

Then, the data matrix was normalized according to min-max normalization approach as it is robust to any distributions of each feature and, it leads to make unitless measurement for each. Now, the matrix data is prepared for applying dimension reduction techniques.

DIMENSION REDUCTION TECHNIQUES FOR THE OBTAINED MATRIX DATA

Supervised learning is a well-known technique, and it is applied when the response variable Y is measured along with a set of p features $X_1, X_2, \ldots, X_p$. But sometimes in practice, there is

only a set of features $X_1, X_2, \ldots, X_p$ measured on n observations, and the interest is not about prediction as there is no an associated response variable Y. The developed technique in this scenario is called unsupervised. There are two types of unsupervised learning: Principial Components Analysis (PCA) and Clustering. The PCA allows us to explain variation of the original set in a smaller set of variables. Assume that in original set we have $X_1, X_2, \ldots, X_p$, in the PCA we transform this set to a smaller set $Z_1, Z_2, \ldots, Z_m$ such that $m < p$ and this new set is a linear combination of the original set by having this assumption that there are orthogonal. For example, the first principal component of the original set of features $X_1, X_2, \ldots, X_p$ is

$Z_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p$, where $a_{11}, a_{21}, \ldots, a_{p1}$ refer to the loadings of the fist principal component. It has been proved that these values are eigenvector of the covariance matrix of $X_1, X_2, \ldots, X_p$. And the eigenvalues of the matrix are the (sample) variances of the principal components. This transformation reshapes the original features and put it in a new direction that has dominant variation. Because the eigenvalues are variances of the principal components, we can speak of "the proportion of variance explained" by the first *k* components: proportion of variance $\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}$, where λ_1 refers to the varication explains by the first component, and so on (Rencher and Christensen, 2012) . We illustrate the use of PCA on the features extracted from tumors of CT images of lung cancer. PCA was performed after standardizing each variable to have mean zero and standard deviation one. Principal components are computed from the correlation matrix, so the total variance is equal to the number of variables, 110, Figure 5.

The PRINCOMP Procedure

Observations	73
Variables	110

	diagnostics_Image.original_Maxim	diagnostics_Mask.original_VoxelN	diagnostics_Mask.original_Volume	original_shape_Elongation_CT	original_shape_Flatness_CT
Mean	0.0931157066	0.000000000	-.0173714640	0.000000000	0.001592433
StD	0.6027917791	1.000000000	0.9956143452	1.000000000	1.006826018

		diagnostics_Image.original_Maxim	diagnostics_Mask.original_VoxelN	diagnosti
diagnostics_Image.original_Maxim	diagnostics_Image.original_Maximum_CT	1.0000	0.1846	
diagnostics_Mask.original_VoxelN	diagnostics_Mask.original_VoxelNum_CT	0.1846	1.0000	
diagnostics_Mask.original_Volume	diagnostics_Mask.original_VolumeNum_CT	-.0860	-.2099	
original_shape_Elongation_CT	original_shape_Elongation_CT	-.1765	0.1596	

Figure 5. Number of Observations and Simple Statistics

Figure 6 displays the eigenvalues. The first principal component accounts for about 34.1% of the total variance, the second principal component accounts for about 23.9%, and the third principal component accounts for about 11.7%. Note that the eigenvalues sum to the total variance.

The eigenvalues indicate that six components provide a good summary of the data: six components account for 87% of the total variance. Subsequent components account for less than 3% each.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	37.5074711	11.2495152	0.3410	0.3410
2	26.2579558	13.3801409	0.2387	0.5797
3	12.8778149	3.2816660	0.1171	0.6968
4	9.5961489	3.9642312	0.0872	0.7840
5	5.6319177	1.8597818	0.0512	0.8352
6	3.7721359	0.7379104	0.0343	0.8695
7	3.0342255	0.5881559	0.0276	0.8971
8	2.4460696	0.4463551	0.0222	0.9193
9	1.9997145	0.6592036	0.0182	0.9375
10	1.3405109	0.3931358	0.0122	0.9497
11	0.9473751	0.0209305	0.0086	0.9583
12	0.9264446	0.1975167	0.0084	0.9667
13	0.7289280	0.3221821	0.0066	0.9733
14	0.4067458	0.0796909	0.0037	0.9770
15	0.3270549	0.0380985	0.0030	0.9800
16	0.2889565	0.0354545	0.0026	0.9826

Figure 6. Results 1 of Principal Component Analysis

Figure 7 displays the eigenvectors. From the eigenvector's matrix, we can represent the first principal component, *Prin1*, as a linear combination of the original variables:

$$Prin1 = 0.090742\ diagnostics_{Image}.original_{Maxim} + 0.066591\ diagnostics_{Mask}.original_{VoxelN} + \cdots - 0.003032\ original\_ngtdm\_Strength.$$

Similarly, the second principal component, *Prin2*, is

$$Prin2 = -.016307\ diagnostics_{Image}.original_{Maxim} + 0.141508\ diagnostics_{Mask}.original_{VoxelN} + \cdots - .123923\ original\_ngtdm\_Strength\ ,$$

where the variables are standardized.

		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
diagnostics_Image.original_Maxim	diagnostics_Image.original_Maximum_CT	0.090742	-.016307	-.079595	0.071713	0.167295	0.018480	0.044174	-.131306
diagnostics_Mask.original_VoxelN	diagnostics_Mask.original_VoxelNum_CT	0.066591	0.141508	0.043505	0.148088	-.054334	-.000869	0.103905	-.027866
diagnostics_Mask.original_Volume	diagnostics_Mask.original_VolumeNum_CT	-.005584	-.029517	-.006751	-.017912	-.087617	0.411841	-.208858	0.035149
original_shape_Elongation_CT	original_shape_Elongation_CT	-.012430	0.036743	-.058551	0.055810	-.026677	-.212245	-.045605	0.123160
original_shape_Flatness_CT	original_shape_Flatness_CT	0.010475	0.013499	-.005282	-.016572	0.022336	0.400037	-.307297	-.119354
original_shape_LeastAxisLength_C	original_shape_LeastAxisLength_CT	0.010475	0.013499	-.005282	-.016572	0.022336	0.400037	-.307297	-.119354
original_shape_MajorAxisLength_C	original_shape_MajorAxisLength_CT	0.064922	0.140009	0.057502	0.082824	-.045241	0.173325	-.058638	-.083686
original_shape_Maximum2DDiameter	original_shape_Maximum2DDiameterColumn_CT	0.063782	0.128595	0.040083	0.074384	-.010871	0.226019	-.084468	-.060202
original_shape_Maximum2DDiamet_1	original_shape_Maximum2DDiameterRow_CT	0.060931	0.138296	0.039181	0.109398	-.051603	0.138859	-.113969	-.105672
original_shape_Maximum2DDiamet_2	original_shape_Maximum2DDiameterSlice_CT	0.041130	0.144533	0.083257	0.049674	-.117150	0.106174	0.118809	0.011667
original_shape_Maximum3DDiameter	original_shape_Maximum3DDiameter_CT	0.042027	0.140081	0.076434	0.041225	-.102688	0.230152	0.011297	-.027978
original_shape_MeshVolume_CT	original_shape_MeshVolume_CT	0.066834	0.141395	0.040922	0.148321	-.050817	-.002871	0.108114	-.021143
original_shape_MinorAxisLength_C	original_shape_MinorAxisLength_CT	0.064248	0.154175	0.035496	0.119905	-.049007	0.028532	0.003589	-.005112
original_shape_Sphericity_CT	original_shape_Sphericity_CT	-.058327	-.160772	-.052340	-.047140	0.031087	0.046153	0.085196	-.041356
original_shape_SurfaceArea_CT	original_shape_SurfaceArea_CT	0.066711	0.141642	0.042320	0.147906	-.053011	0.000328	0.104837	-.024992
original_shape_SurfaceVolumeRati	original_shape_SurfaceVolumeRatio_CT	-.032689	-.137451	-.012810	-.013283	-.016645	0.170424	0.117024	-.094562
original_shape_VoxelVolume_CT	original_shape_VoxelVolume_CT	0.066920	0.141710	0.041152	0.148020	-.051322	0.000064	0.105358	-.022322
original_firstorder_10Percentile	original_firstorder_10Percentile_CT	-.055902	0.119793	-.153021	-.014403	0.157773	0.049393	0.054312	-.011030
original_firstorder_90Percentile	original_firstorder_90Percentile_CT	0.056515	0.061102	-.037497	0.000782	0.301498	0.010222	0.099460	-.217470
original_firstorder_Energy_CT	original_firstorder_Energy_CT	0.076086	0.076698	0.121192	0.135107	-.139506	-.049601	-.015028	-.099402
original_firstorder_Entropy_CT	original_firstorder_Entropy_CT	0.149938	-.063165	0.027916	0.012335	-.026610	-.012554	-.029926	0.052764
original_firstorder_Interquartil	original_firstorder_InterquartileRange_CT	0.074727	-.124294	0.111503	0.047225	0.032634	0.051911	0.106495	0.013372
original_firstorder_Kurtosis_CT	original_firstorder_Kurtosis_CT	-.017315	0.122718	-.008432	-.172347	0.015091	0.069886	0.054738	0.080363
original_firstorder_Maximum_CT	original_firstorder_Maximum_CT	0.081884	0.044109	-.005824	0.033415	0.118976	0.074041	0.074851	0.189701
original_firstorder_MeanAbsolute	original_firstorder_MeanAbsoluteDeviation_CT	0.098774	-.105018	0.156932	-.003761	0.002617	-.008715	0.021161	-.063078
original_firstorder_Mean_CT	original_firstorder_Mean_CT	-.013109	0.115528	-.096706	-.021163	0.249758	-.011749	0.056355	-.204266

Figure 7. Results 2 of Principal Component Analysis

The first component is a measure of the overall nature of tumor because the first two eigenvector shows approximately equal loadings on all variables. The third eigenvector has high positive loadings on the variables *diagnsitcs_Msak.original_VoxelN* and high negative loadings on the variables *Original_Shape_Sphericity*. This component seems to measure the shape of tumor compared to first order statistic. The interpretation of the third component is not obvious.

Also, we can illustrate the pairwise component score plots for the first three components, with a 95% prediction ellipse overlaid on each scatter plot. Figure 8 shows the plot of the first three components. In the left and right panels, you can identify regional trends in the plot of the first two components. Assuming that the first two components are from a bivariate normal distribution, the ellipse identifies extracted features for patients 20, 37, and 72 as a possible outlier.

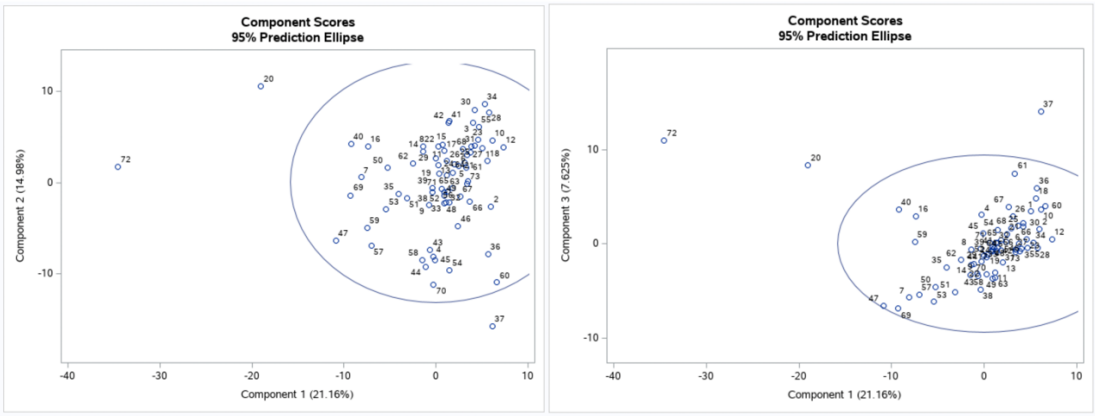


Figure 8. Plot of the First Three Component Scores

In every application, a decision must be made on how many principal components should be retained in order to effectively summarize the data. Scree plot is one of techniques that people often use to determine how many principal components is needed to keep. This is a plot of λ_i versus i, and look for a natural break between the "large" eigenvalues and the "small" eigenvalues. Figure 9 shows that the first six eigenvalues form a steep curve, followed by a bend and then a straight-line trend with shallow slope. The recommendation is to retain those eigenvalues in the steep curve *before* the first one on the straight line. Thus, in Figure 9 (Left panel), six components would be retained. Additionally, the right panel of Figure 9 confirms that six components explain enough number of variations from original data which is about 87% in total.

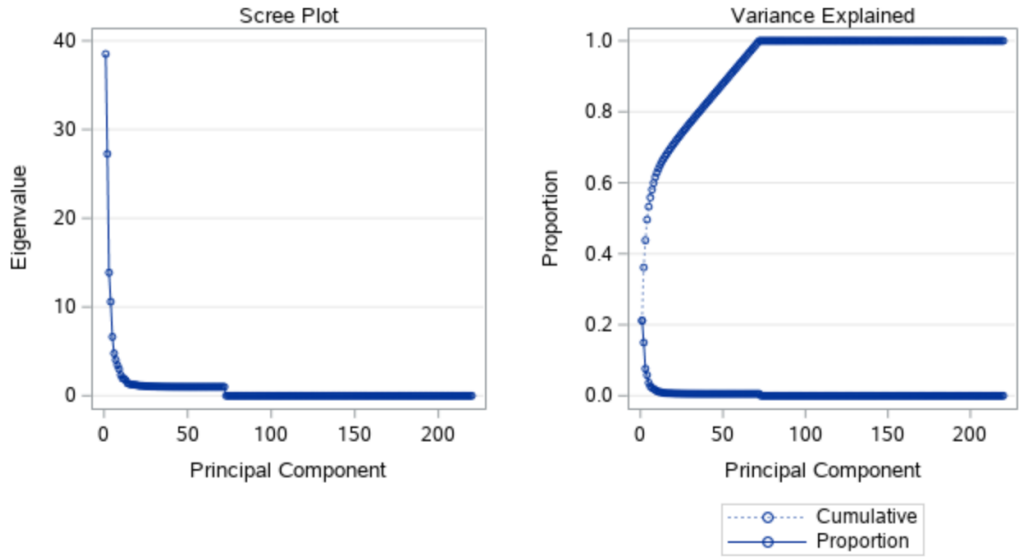


Figure 9. Scree graph for eigenvalues of lung cancer data (Left panel), Plot of variance explained by principal components (Right panel)

The second techniques of dimension reduction that we would explain in this paper is clustering. In clustering the goal is to find homogeneous subgroups among the observations. There are two well-known clustering techniques: K-means and hierarchical. In K-means we do clustering by having a pre-specified number of clusters. However, in hierarchical, we do not know how many clusters we want. The K-means clustering results from a fundamental mathematical idea. Assume that $C_1, C_2, \ldots, C_K$ represents sets including the observations clustered into K subgroups of the original data. These sets meet two properties (James et al., 2013):

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. It means the union of all clusters lead to the whole observations.
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. It means clusters are pairwise mutually exclusive.

Now, we would like to state the algorithm behind K-means clustering techniques.

1. Randomly assign a number to each observation from 1 to K. This calls an initial clustering for the observations.
2. Repeat the following process till the cluster assignments stop changing.
    a. For each of the K clusters, calculate the kth cluster centroid which is the vector of the p feature means for the observations in the kth cluster.

b. Use Euclidean distance for assigning each observation to the nearest centroid.

On the other hand, in hierarchical clustering there is not a requirement of knowing a particular choice of K. One of the advantages of this approach is that the obtained clusters can be visualized through a plot called dendrogram, which is a tree-based representation of the observations. The algorithm of hierarchical is as follows (James et al., 2013):

1. Start with n observations and compute Euclidean distance of all $\frac{n(n-1)}{2}$ for all pairwise dissimilarities. Consider each observation as an independent cluster.
2. Repeat the following process for $i = n, n-1, \ldots, 2$:
   a. Test all pairwise of each dissimilar cluster and determine the pair of clusters that are most similar. Connect these two clusters.
   b. Compute the new pairwise dissimilarities for the rest of $i-1$ clusters.

In this algorithm, the concept of dissimilarity between a pair of observations is extended to a pair of groups of observations, which defined as a linkage. There are four common types of linkage: complete, average, single (Ward's), and centroid. The summary of these linkages as follows (James et al., 2013):

1. Complete: In this approach, all pairwise dissimilarities between the observations in the clusters are computed and the maximum one will be recorded.
2. Single (Ward's): In this method, all pairwise dissimilarities between the clusters are computed and the minimum one will be recorded.
3. Average: In this approach, all pairwise dissimilarities between the clusters are computed and the average of dissimilarities will be recorded.
4. Centroid: In this technique, the dissimilarities between the mean vector of for cluster A (centroid) and the mean vector of for cluster B (centroid) are computed.

Now, we examine the hierarchical approach on the extracted features from lung cancer data. First, we specify that approximately 3% of the pairs are included in the estimation within-cluster covariance matrix, and also, Ward's minimum-variance clustering method is used. The results of cluster history are summarized in Figure 10. This displays the last 15 generations of the cluster history.

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Tie
15	CL29	CL33	6	0.0044	.960	.	.	99.0	10.2	
14	CL24	CL20	20	0.0059	.954	.821	20.7	93.9	15.5	
13	CL14	CL18	25	0.0092	.945	.808	19.4	85.5	12.3	
12	CL27	OB18	4	0.0094	.935	.794	18.5	80.2	17.4	
11	CL23	CL17	11	0.0111	.924	.777	17.6	75.6	15.7	
10	CL16	CL13	40	0.0221	.902	.759	15.1	64.5	25.2	
9	CL15	CL11	17	0.0230	.879	.738	13.4	58.2	14.6	
8	OB37	OB61	2	0.0246	.855	.713	12.3	54.6	.	
7	CL12	CL9	21	0.0475	.807	.684	9.36	46.0	15.8	
6	CL21	CL8	8	0.0494	.758	.649	7.53	41.9	9.9	
5	CL10	CL19	42	0.0559	.702	.604	6.28	40.0	38.3	
4	CL5	CL6	50	0.1536	.548	.535	0.61	27.9	38.1	
3	CL4	OB20	51	0.1662	.382	.400	-.75	21.6	23.5	
2	CL7	CL3	72	0.1747	.207	.230	-1.0	18.6	19.8	
1	CL2	OB55	73	0.2072	.000	.000	0.00	.	18.6	

Figure 10. Cluster History

First listed are the number of clusters and the names of the clusters joined. The observations are identified either by the ID value or by CLn, where n is the number of the cluster. Next, Figure 10 displays the number of observations in the new cluster and the semi-partial R square. The latter value represents the decrease in the proportion of variance accounted by joining the two clusters. Next listed is the squared multiple correlation, R square, which is the proportion of variance accounted for by the clusters. Figure 10 shows that, when the data are grouped into six clusters, the proportion of variance accounted for by clusters (R square) is about 75%. The approximate expected value of R square is given in the ERSq column. This expectation is approximated under null hypothesis that the data have a uniform distribution instead of forming distinct clusters. The next three columns display the values of the cubic clustering criterion (CCC), pseudo F (PSF), and t^2 (PST2) statistics. These statistics are useful for estimating the number of clusters in the data. The final column in Figure 10 lists ties for minimum distance; a blank value indicates the absence of a tie. A tie means that the clusters are indeterminate and that changing the order of the observations might change the clusters.

Figure 11 plots the tree statistics for estimating the number of clusters. Peaks in the plot of cubic clustering criterion with values greater than 2 or 3 indicate good clusters; peaks with values between 0 and 5 indicate possible clusters. Large negative values of the CCC can indicate outliers. In Figure 11, there is a local peak of the CCC when the number of clusters is three. The CCC drops at four clusters and then steadily increases.

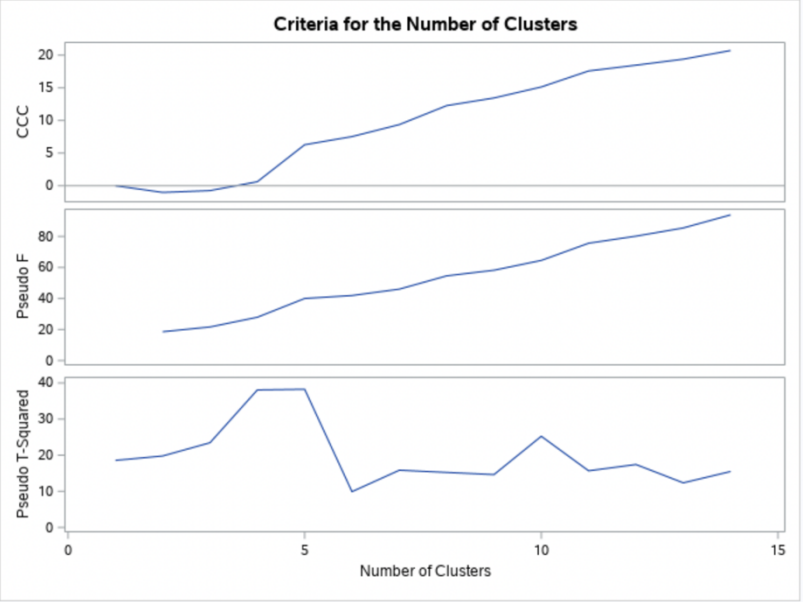


Figure 11. Plot of Statistics for Estimating the Number of Clusters

Another method of judging the number of clusters in a data set is to look at the pseudo F statistic (PFS). Relatively large values indicate good number of clusters. In Figure 11, the pseudo F statistic suggests five or more up to thirteen.

To interpret the values of the pseudo t^2 statistic, look down the column or look at the plot from right to left until you find the first value that is markedly larger than the previous value, then move back up the column or to the right in the plot by one step in the cluster history. In Figure 11, we can see possibly good clustering levels at six clusters, eight clusters, ten clusters, and eleven clusters.

Considered together these statistics suggest that the data can be clustered into six clusters.

Figure 12 displays the dendrogram. The figure provides a graphical view of the information

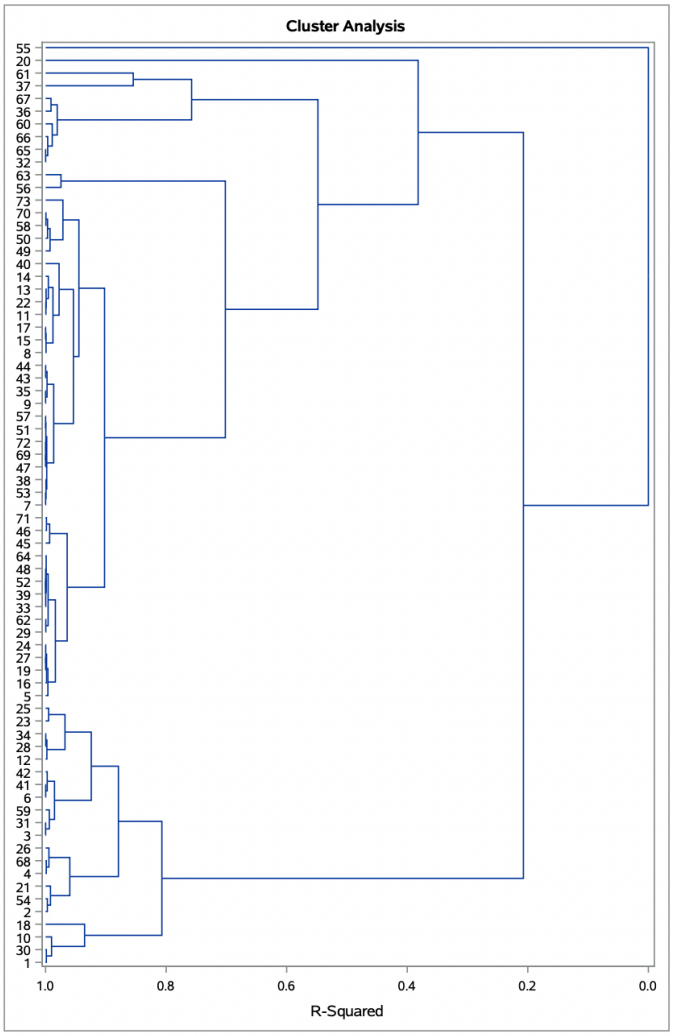


Figure 12. Dendrogram of Clusters versus R-Square Values

-in Figure 12. As the number of branches grow to the left from the root, the R square approaches 1; the first six clusters (branches of the tree) account for over half of the variation (about 75%, from Figure 12). In other words, only six clusters are enough to explain over three-fourths of the variation.

CONCLUSION

In this paper, we have shown how PCA and Clustering in the area of healthcare especially in cancer medical images. Dealing with dimensional of big data is very challenging because the data contain a large set of variables and features. Therefore, dimension reduction techniques are very important task that a data scientist needs to do before conducting any statistical analysis. Several computational techniques have been developed in order to do dimension reductions, but the ones that we discussed were PCA and Clustering. PCA

suggested that about six components would explain enough number of variations from original data which is about 87% in total. On the other hand, Clustering discovered about six similar subgroups in data. In other words, the first six clusters account for over half of the variation about 75%. We can conclude that both approaches suggested we can categorize the data into six subcategories to adhere with the dominate variation in the data. Consequently, if a researcher wants to implement predicting models and supervised techniques, he or she can process the analysis using the six discovered components.

DISCUSION

The PCA is a linear combination of features in original data. So, if the structure in the original data doesn't follow linearity, PCA is not a good representative of original data and cannot capture the amount number of variations. Therefore, non-linear techniques such as manifold learning should apply as it can be adjustable with the non-linearity nature of the data. On the other hands, there are some practical issues with clustering techniques. In the case of hierarchical clustering what kind of dissimilarity measurements should be used is questionable. There is the same issue about choosing the linkage. Also, what initial value should be considered to put a cut point in the dendrogram to get clusters is still under developing. In case of K-means clustering, the determination of optimal cluster is challenging. Consequently, in the future research we could use other methods in machine learning such as Local fisher's discriminate analysis (LFDA), Canonical correlation analysis (CCA), Non-negative matrix factorization (NMF), and Manifold learning-based algorithm for addressing our scientific hypothesis from another point of view.

REFERENCES

Afshar, P., Mohammadi, A., Plataniotis, K. N., Oikonomou, A., & Benali, H. (2019). From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. *IEEE Signal Processing Magazine*, *36*(4), 132-160.

Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, Andre´ Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

Philippe Lambin, Steven F Petit, Hugo JWL Aerts, Wouter JC van Elmpt, Cary JG Oberije, Maud HW Starmans, Ruud GPM van Stiphout, Guus AMS van Dongen, Kristoff Muylle, Patrick Flamen, et al. The estro breur lecture 2009. from population to voxel-based radiotherapy: exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer. *Radiotherapy and oncology*, 96(2):145–152, 2010.

Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, Rene´ Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):1–9, 2014.

Petros-Pavlos Ypsilantis, Musib Siddique, Hyon-Mok Sohn, Andrew Davies, Gary Cook, Vicky Goh, and Giovanni Montana. Predicting response to neoad- juvant chemotherapy with pet imaging using convolutional neural networks. *PloS one*, 10(9):e0137036, 2015.

Issam El Naqa. The role of quantitative pet in predicting cancer treatment outcomes. *Clinical and translational imaging*, 2(4):305–320, 2014.

WP Wu, J Wang, L Lu, B Xie, Y Wu, and A Kumar. Syntheses, hirshfeld surface analyses, and luminescence of four new complexes. *Russian Journal of Coordination Chemistry*, 42(1):71–80, 2016.

Chintan Parmar, Patrick Grossmann, Derek Rietveld, Michelle M Rietbergen, Philippe Lambin, and Hugo JWL Aerts. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Frontiers in oncology*, 5:272, 2015.

AndreasBannach,Ju ̈rgenBernard,FlorianJung,Jo ̈rnKohlhammer,Thorsten May, Kathrin Scheckenbach, and Stefan Wesarg. Visual analytics for radiomics: Combining medical imaging with patient data for clinical research. In *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*, pages 84–91. IEEE, 2017.

Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):1–7, 2018.

Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523, 2013.

Qing Feng, Meilei Jiang, Jan Hannig, and JS Marron. Angle-based joint and individual variation explained. *Journal of multivariate analysis*, 166:241–265, 2018.

Michael J O'Connell and Eric F Lock. R. jive for exploration of multi-source molecular data. *Bioinformatics*, 32(18):2877–2879, 2016.

Guoxu Zhou, Andrzej Cichocki, Yu Zhang, and Danilo P Mandic. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE transactions on neural networks and learning systems*, 27(11):2426– 2439, 2015.

Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion- Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

Li, P., Wang, S., Li, T., Lu, J., HuangFu, Y., & Wang, D. (2020). *A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis* [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2020.NNC2-0461

Parekh, V., & Jacobs, M. A. (2016). Radiomics: a new application from established techniques. *Expert review of precision medicine and drug development*, *1*(2), 207-226.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis*, John Wiley & Sons, Inc.

Tian, J., Dong, D., Liu, Z., & Wei, J. (2021). *Radiomics and Its Clinical Application: Artificial Intelligence and Medical Big Data*. Academic Press.

Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jen- nings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012.

Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*, *54*(5), 3473-3515.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

<Mostafa Zahed>
<zahedm@etsu.edu>

<Maryam Skafyan>
<skaf9868@bears.unco.edu >

APPEDIX

<SAS CODE>

```
/* Generated Code (IMPORT) */
/* Source File: MatrixData_Lung_CT_1_74_Modify.xlsx */
/* Source Path: /home/mostafazahed0/SESUG */
/* Code generated on: 7/17/22, 6:51 PM */

%web_drop_table(WORK.IMPORT);
```

```
FILENAME REFFILE '/home/mostafazahed0/SESUG/Lung_Cancer_CT.xlsx';

PROC IMPORT DATAFILE=REFFILE
      DBMS=XLSX
      OUT=Lung;
      GETNAMES=YES;
RUN;

PROC CONTENTS DATA=Lung;

RUN;

%web_open_table(WORK.IMPORT);


/*view mean and standard deviation of dataset*/
proc means data=Lung Mean StdDev ndec=3;
run;

/*normalize the dataset*/
proc stdize data=Lung out=normalized_Lung;
   *var values;
run;

/*print normalized dataset*/
proc print data=normalized_Lung;

/*view mean and standard deviation of normalized dataset*/
proc means data=normalized_Lung Mean StdDev ndec=2;
   *var values;
run;


proc factor data=normalized_Lung simple corr score scree;
run;
```

```
proc princomp out=normalized_Lung plots= score(ellipse ncomp=6);
run;



/* Perfoming Cluster Analysis */
ods graphics on;
proc cluster data=normalized_Lung method = centroid ccc print=15
outtree=Tree;
*var diagnostics_Image.original_Maxim--original_ngtdm_Strength_CT;
run;
ods graphics off;

/* Retaining 9 clusters */
proc tree data=Tree noprint ncl=8 out=out;
*copy diagnostics_Image.original_Maxim--original_ngtdm_Strength_CT;
run;

proc print data=out;
run;

/* To create a Scatterplot */
proc candisc out=;
class cluster;
var diagnostics_Image.original_Maxim original_ngtdm_Strength_CT;
*var ALL;
run;

proc sgplot data =Tree;
title "Cluster Analysis for Lung datasets";
scatter y = can2 x = can1 / group = cluster;
run;



/*Run the fastclus multiple times*/
%macro kmean(K);

proc fastclus data=normalized_Lung out=outdata&K. maxclusters= &K.
maxiter=100 converge=0;
```

```
var v1-v4;
run;

* Canonical Discriminant Analysis;
proc candisc data=outdata4 out=egclustcan;
class cluster;
var v1-v4;
run;

/*Plots the two canonical variables generated from PROC CANDISC, can1 and
can2*/

proc sgplot data=egclustcan;
scatter y=can2 x=can1 / group=cluster;
run;

proc print data=Lung;
run;

/*******************************/
OPTIONS NONUMBER NODATE LS=95;
ODS pdf FILE="/home/mostafazahed0/SESUG/ClusteringSASoutput.pdf" notoc ; **
Sets the output to print to pdf ;

ODS NOPTITLE; ** Don't print "The REG Procedure"-type titles;

proc aceclus data=Lung out=Ace p=.03 noprint;
run;

proc print data=Ace;
run;

ods graphics on;
proc cluster data=Ace method=ward ccc pseudo print=15 out=tree
plots=den(height=rsq);
var can1-can110;
*id Obs;
run;
ods graphics off;
```

```
ODS pdf CLOSE;

proc tree data=Tree out=New nclusters=3 noprint;
height _rsq_;
copy can1-can2;
*id country;
run;

proc sgplot data=New;
scatter y=can2 x=can1 / group=cluster;
run;

/*************************/
ods graphics on;
title 'Using METHOD=AVERAGE';
proc cluster data=Ace method=average pseudo ccc pseudo print=15 out=tree
plots=den(height=rsq);
var can1-can110;
*id Obs;
run;
ods graphics off;

proc tree data=Tree out=New nclusters=5 noprint;
height _rsq_;
copy can1-can2;
*id country;
run;

proc sgplot data=New;
scatter y=can2 x=can1 / group=cluster;
run;

/*************************/
ods graphics on;
title 'Using METHOD=CENTROID';
proc cluster data=Ace method=centroid pseudo ccc pseudo print=15 out=tree
plots=den(height=rsq);
```

```
var can1-can110;
*id Obs;
run;
ods graphics off;

proc tree data=Tree out=New nclusters=5 noprint;
height _rsq_;
copy can1-can2;
*id country;
run;

proc sgplot data=New;
scatter y=can2 x=can1 / group=cluster;
run;

/*************************/
ods graphics on;
title 'Using METHOD=DENSITY K=3';
proc cluster data=Ace method=density k=3 pseudo ccc pseudo print=15 out=tree;
*plots=den(height=rsq);
var can1-can110;
*id Obs;
run;
ods graphics off;

proc tree data=Tree out=New nclusters=5 noprint;
height _rsq_;
copy can1-can2;
*id country;
run;

proc sgplot data=New;
scatter y=can2 x=can1 / group=cluster;
run;

/*************************/
ods graphics on;
title 'Using METHOD=SINGLE';
```

```
proc cluster data=Ace method=single pseudo ccc pseudo print=15 out=tree;
*plots=den(height=rsq);
var can1-can110;
*id Obs;
run;
ods graphics off;

proc tree data=Tree out=New nclusters=5 noprint;
height _rsq_;
copy can1-can2;
*id country;
run;

proc sgplot data=New;
scatter y=can2 x=can1 / group=cluster;
run;

/*************************/
ods graphics on;
title 'Using METHOD=TWOSTAGE K=3';
proc cluster data=Ace method=twostage k=3 pseudo ccc pseudo out=tree
plots=den(height=rsq);
var can1-can110;
*id Obs;
run;
ods graphics off;

proc tree data=Tree out=New nclusters=5 noprint;
height _rsq_;
copy can1-can2;
*id country;
run;

proc sgplot data=New;
scatter y=can2 x=can1 / group=cluster;
run;

/*************************/
```

```
ods graphics on;
title 'Using METHOD=Ward';
proc cluster data=Ace method=ward pseudo ccc pseudo out=tree
plots=den(height=rsq);
var can1-can110;
*id Obs;
run;
ods graphics off;

proc tree data=Tree out=New nclusters=5 noprint;
height _rsq_;
copy can1-can2;
*id country;
run;

proc sgplot data=New;
scatter y=can2 x=can1 / group=cluster;
run;
```