SESUG Paper 205-2023

Unsupervised Dimension Reduction Techniques for Lung Cancer Diagnosis Based on Radiomics

Janet Akoth Kireta and Mostafa Zahed, Mathematics and Statistics Department, East Tennessee State University, TN, USA.

ABSTRACT

Over the years, cancer has increasingly become a global health problem. For successful treatment, early detection and diagnosis are critical. Radiomics is the use of CT, PET, MRI, or Ultrasound imaging as input data, extracting features from image-based data, and then using machine learning for quantitative analysis and disease prediction. Feature reduction is critical as most quantitative features can have unnecessary redundant characteristics. This research aims to use machine learning techniques to reduce the number of dimensions, thereby rendering the data manageable. Radiomics steps include Imaging, segmentation, feature extraction, and analysis. For this research, large-scale CT data for Lung cancer diagnosis collected by scholars from Medical University in China is used to illustrate the dimension reduction techniques via SAS and Python. The data is available on The Cancer Imaging Archive (TCIA). PyRadiomics through 3D Slicer medical software was used to extract 110 features for 74 out of 130 patients. This research's proposed reduction and analysis techniques entailed; Principal Component Analysis and Clustering analysis (Hierarchical Clustering and K-means). To achieve results for the analyses SAS codes PRINCOMP, CLUSTER, and, FASTCLUS were used. These techniques were equally augmented by computing threshold values and using them to filter out the most salient features using the R program. For the PCA the eigenvalues indicated that three principal components provided a good summary of the data accounting for 98.10% of the total variance. The number of features selected was 39, of which 4 were intensity, 7 were shaped, and 28 were texture-based. For clustering analysis, the agglomerative hierarchical clustering algorithm clustered the features into 3 clusters, 21 features were selected whereby 3 were intensity, 3 were shaped and 15 were texture-based features. K-means clustering algorithm with an initial cluster optimum cluster of 3, selected 21 features, of which 4 were intensity, 1 shape, and 15 texture-based features. Overall, all the analyses clearly outlined texture-based features as the most salient category of features.

Keywords: Radiomics, Segmentation, Dimension Reduction, Features Extraction, Feature Selection.

INTRODUCTION

The field of radiomics has rapidly emerged as an important and influential area of contemporary cancer research. It offers a range of potential benefits, particularly in standardizing the analysis of complex imaging data, which ultimately allows for comparative studies across multiple patients and investigations [2]. Identifying key imaging biomarkers through radiomics can significantly improve the accuracy of cancer diagnosis and staging, which can have life-saving implications for patients. Furthermore, the quantitative information that radiomics extracts from images can offer valuable insights into the underlying biology of a tumor, providing clues as to its aggressiveness or how it might respond to different treatments [9]. This information, in turn, can be used to develop tailored treatment plans for patients, identifying those most likely to benefit from specific therapies and those at a greater risk for recurrence or progression. The non-invasive nature of radiomics offers distinct advantages in reducing the need for invasive procedures and enhancing the efficiency of clinical trials [17]. Different types of non-invasive imaging include Molecular imaging which allows clinicians to not only see where a tumor is located in the body but also visualizes the expression and activity of specific molecules (e.g., proteases and protein kinases) and biological processes (e.g., apoptosis, angiogenesis, and metastasis) that influence tumor behavior and/or response to therapy, Anatomical imaging enables the detection of a phenotypic(physical expression of DNA(Deoxyribonucleic Acid)) alteration that is sometimes, but not invariably, associated with cancer, and finally, functional imaging used to study tumor physiology, probe tumor molecular processes, and study tumor molecules and metabolites in vitro and in vivo. These attributes make radiomics an exciting and promising field poised to contribute significantly to advancing cancer research and treatment. Radiomics often encompasses the extraction and analysis of quantitative features from medical images, including but not limited to CT and PET scans. By evaluating tumor size, shape, texture, and density, radiomics offers a promising avenue for advancing personalized medicine [1]. CT and PET scans are widely employed in medical imaging techniques that play an essential role in diagnosing and monitoring cancer. While similar in that they are both non-invasive, the two methods differ in how they generate images. CT scans use X-rays to create detailed, cross-sectional images of internal organs and structures, which can help doctors identify the location and size of tumors. On the other hand, PET scans involve injecting a small amount of radioactive material into the body, which is then used to produce images that reveal the functional activity of tissues (John Hopkins Medicine, 2021). Doctors can analyze these images to assess how cancer cells metabolize nutrients, grow, and spread. Together, these two imaging techniques provide a comprehensive way to monitor cancer without requiring invasive procedures. One critical step in the radiomics workflow is feature extraction, which involves identifying and quantifying the various characteristics of tumors. To accomplish this, segmentation is typically performed to isolate the tumor region, and then multiple methods are used to extract features based on tumor intensity, texture, and shape [18]. Dimension reduction techniques, such as PCA and clustering, are often used to help process and analyze these features. These techniques help to simplify the data by reducing the number of variables and identifying key patterns. More advanced methods have been developed for dimension reduction, such as contrastive Principal Component Analysis (cPCA) and Joint and Individual Variation Explained (JIVE). The cPCA approach can identify low-dimensional structures unique to a particular data set by comparing them to a reference data set. On the other hand, JIVE decomposes variation across multiple data types into joint and individual components [13]. Both methods can help analyze complex medical imaging data. Some of the software tools used for feature extraction include PyRadiomics, 3D Slicer, LIFEx, IBEX, QIFE, and RayPlus. Each device has strengths and limitations, so researchers must carefully consider which best meets their needs.

Overall, the implications of radiomics as a field of study are substantial, particularly as they pertain to diagnosing, treating, and monitoring cancer. By utilizing quantitative data extraction methods from medical images, radiomics can allow researchers to discern patterns in tumor biology that might otherwise remain obscured. This may help shed light on various aspects of a tumor's behavior, such as its aggressiveness or responsiveness to different treatment modalities. As such, radiomics has the potential to contribute significantly to our overall understanding of cancer and to facilitate the development of more effective and personalized therapies.

The goal of studying cancer is to develop safe and effective methods to prevent, detect, diagnose, treat, and, ultimately, cure the collections of diseases we call cancer. The better we understand this disease, the more progress we will make toward diminishing the tremendous human and economic tolls of cancer. Recent advances in medical imaging, such as radiomics, have shown great potential in this regard. Radiomics allows for the extraction and analysis of large data sets from imaging techniques such as CT and PET scans. This, in turn, provides a more comprehensive understanding of tumor growth and development. As such, using radiomics in cancer detection and analysis represents a promising avenue for future research, potentially leading to significant improvements in diagnosis, treatment, and patient outcomes. The process may however turn out to be very hectic given the features obtained from radiological images are so immense. Therefore, there is a dire need to have the data matrix in its simplest form to give way for prognosis, therapy, and any other objective such kinds of research would intend to accomplish. To accomplish this, the research intended to answer the following questions;

RQ 1. Is there a way to reduce the number of variables from 110 to a lesser number that would make the process of working with the data simple?

RQ 2. Is any of the feature categories most significant for our analysis?

Overall, the ultimate intention of the analysis would be to generate a data matrix with fewer and very significant features that can be used in the future as new predictor variables to do predictions on the Lung Cancer data.

RESEARCH METHODOLOGY

The techniques used to address the research question included data description and analysis techniques.

DATA DESCRIPTION

The data was collected by Huiping Han, Funing Yang, and Rui Wang of Harbin from the Medical University in Harbin in China [18]. This data is available on The Cancer Imaging Archive (TCIA). The workflow of radiomics includes; medical imaging, segmentation of the tumor region, feature extraction based on intensity, texture, and shape [11], finally, analysis of the features, Figure 1.

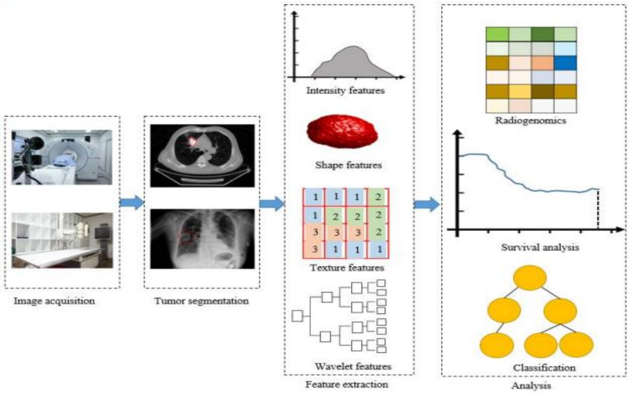


Figure 1: The Radiomics Workflow

Source: https://wiki.cancerimagingarchive.net

This dataset consists of CT DICOM images of 130 patients with lung cancer. The XML Annotation files which include the location of the tumor were provided by five academic radiologists with high expertise in lung cancer. To visualize the annotation boxes on the tumor of the DICOM images [11], python codes through the terminal were used to pull out the images and put the location of tumor in a box, Figure 2.

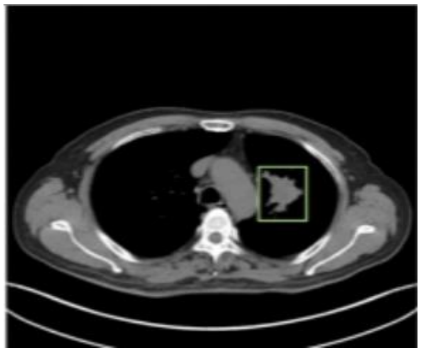


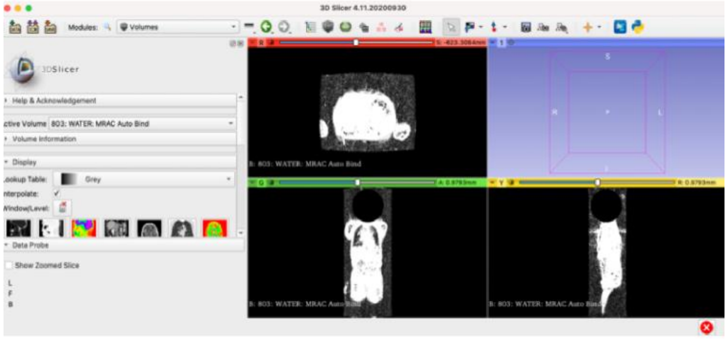
Figure 2:Visualization of the annotation box on the CT-DICOM images.

Source: https://wiki.cancerimagingarchive.net

The process of data acquisition is a outlined;

a) Software tools for extracting features:

There are massive software tools available for extracting tumor features from medical images. Some standard options include PyRadiomics, 3D Slicer, LIFEx, IBEX, QIFE, and RayPlus. Each of the devices has its drawbacks and advantages. It is therefore at the researcher's discretion to identify which best aligns with his intended objectives. For example, PyRadiomics is a flexible open-source platform capable of extracting a wide array of features, but it requires some programming knowledge in Python [9]. 3D Slicer, on the other hand, is a free and open-source application designed to facilitate the development of new functionality in 3D Slicer extensions [5], Figure 3. LIFEx is another option that offers a user-friendly interface and powerful features for tumor segmentation, feature extraction, and radiomics analysis. Ultimately, the choice of software tool depends on the researcher's goals and expertise.



3D Slicer 4.11.20200930

Modules: Volumes

ctive Volume 803: WATER: MRAC Auto Bind

Lookup Table: Grey

nterpolate: ✓

□ Show Zoomed Slice

Figure 3:Loading Lung-CT-PET Images.

Source: https://wiki.cancerimagingarchive.net

b) Extracting features from CT medical images of lung cancer.

Features that are extracted can be generally classified into three main categories [2]: First-order radiomics which has Intensity-based features and Shape based features, second-order radiomics which has Texture-based features extracted based on different descriptive matrices (Gray level co-occurrence matrix (GLCM), Gray level run length matrix (GLRLM), Neighborhood gray-tone difference matrix (NGTDM), Gray level zone length matrix (GLZLM), Figure 4.

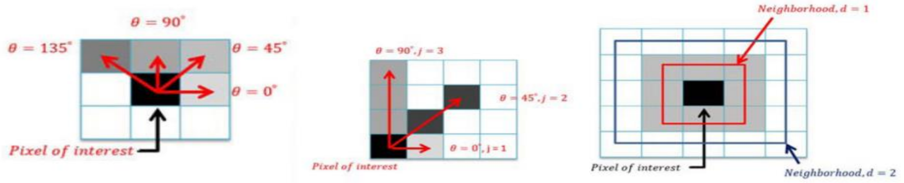


Figure 4. Texture Features from left to right: GLCM (gray level co-occurrence matrix), GLRLM (gray level run length matrix) and NGTDM (neighborhood gray tone difference matrix) (Parekh and Jacobs (2016))

Figure 4: Texture Features (Parekh and Jacobs (2016)).

The last category, higher-order radiomics applies the use of filters to extract features from images through Wavelet which decomposes tumor images into different frequency domains (such as horizontal, vertical, and diagonal) and then extracts the tumor shape, intensity, texture, and other information. Fourier features capture gradient information while Minkowski Functional (MF) is a common higher-order feature extractor considering the patterns of pixels with intensities above a predefined threshold.

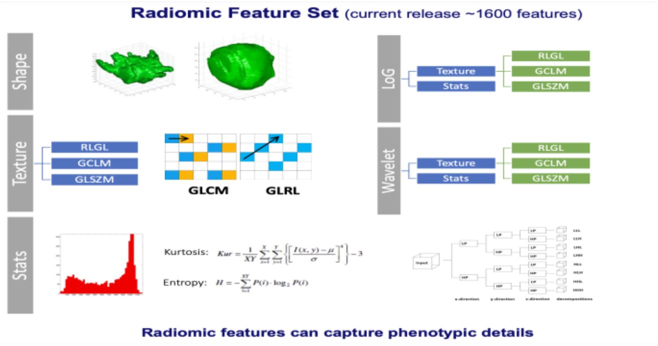


Figure 5: Categories of Features

Source: https://wiki.cancerimagingarchive.net.

c) Extraction process: Out of the 130 patients under consideration, the extraction of features was done on 74 patients because the provided annotation files did not work for all 130 patients. A 3D slicer was used to do the segmentation process as indicated by the yellow circle around the tumor, Figure 6.

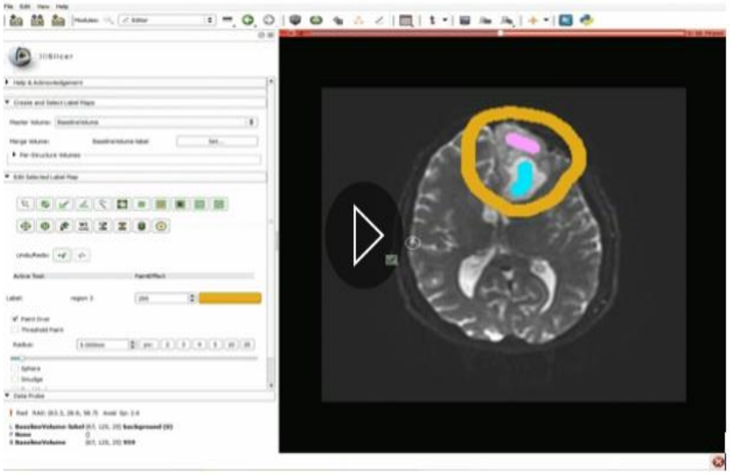


Figure 6: A 3D slicer segmenting the Tumor

The PyRadiomics package is available in the 3D slicer was then used to extract features from the tumor segmentations for all patients, Figure 7.

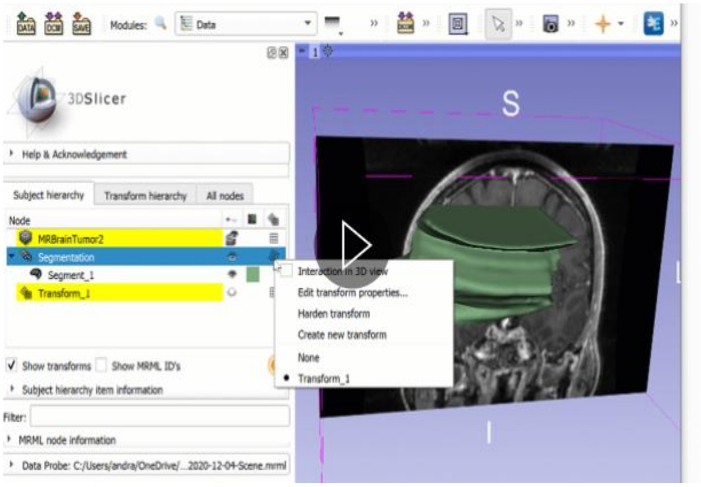


Figure 7: PyRadiomics package extracting features.

d) Resulting Data Matrix

After the whole process of extraction, an unsupervised data matrix was obtained with a dimension of 74 by 110. Each row represented the patient, and each column was for the extracted feature depending on the categories earlier discussed. The 110 features acquired were quantitative variables. Finally, the data matrix was normalized according to the min-max normalization approach as it is robust to any feature distributions and leads to making unitless measurements for each feature [11]. The features on the columns were renamed since the original names were too long to enable data visualization through graphs.The column names range from diagnostics_Image.original_Maximum_CT,…, original_ngtdm_Strength_CT were renamed to $V_1, \ldots, V_{110}$. Since the resulting matrix has 74 rows by 110 columns, most reduction techniques algorithms such as PCA, hierarchical and even k-means clustering cannot handle such a data format successfully, it is for this reason that the normalized data set was transformed into a square correlation matrix such that the new dimension was 108 by 108. It is after this transformation the data matrix was finally ready for applying dimension reduction techniques.

Analysis Techniques

As dictated by the research objective, dimension-reduction techniques are applied to render the data more manageable. These approaches included feature extraction and selection [16]. Feature extraction techniques are further categorized into; supervised and unsupervised learning. Supervised learning is a technique that considers the relation of features with class labels and features are selected mostly based on their contribution to distinguish classes, while, unsupervised learning does not consider the class labels and its objective is to remove redundant features [3]. Because the obtained data matrix is unsupervised, therefore a further linear exploration into the classification of unsupervised learning techniques Principal Component Analysis (PCA) was done to transform the original data into a new set of features that retain most of the original dataset's information. Selecting the appropriate dimension reduction technique is a function of the specific dataset and research objectives. Employing these techniques allows researchers to improve computational efficiency, avoid the curse of dimensionality, and pinpoint the most salient features in the dataset.

Reduction Technique in Radiomics

1. Principal Component Analysis (PCA)

A feature transformation technique that reduces the correlation between sampled variables [1 4] say $x_1, x_2, \ldots, x_p$. Using an orthogonal transformation, PCA generates new variables referred to as principal components $Pc_1, Pc_2, \ldots, Pc_m$ that retains many of the properties of the original variables given $m < p$. This approach enables the creation of various features through linear combinations of the main components, which maximize variance and improve predictability [6], Figure 8.

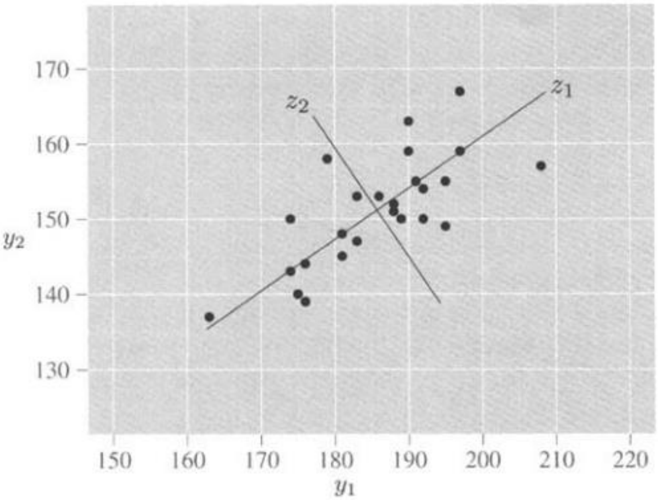


Figure 8: Principal Component Analysis
Source: (Rencher and Christensen (2012)).

There are five main steps to conducting PCA:

(a) Standardize the data: Calculate the mean of all the dimensions of the data set, except the labels. Scale the data so that each variable contributes equally to the analysis. In the equation given below, z is the scaled value, x is the initial, and μ and σ are the mean and standard deviation, respectively.

$$Z = \frac{x - \mu}{\sigma},$$

(b) Compute the covariance matrix: Identifying highly correlated variables is a crucial step in data analysis. These variables often contain redundant information, which can hinder the accuracy of statistical models and analyses. Utilizing a covariance matrix allows for the examination of correlations between all possible variable pairs within a given data set and
facilitates the removal of any superfluous variables.

$$Cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}),$$

where x is the mean of the predictor variables, y is the mean of the response variables, n is the sample size and i refers to each observation.
Basing the PCA on the covariance matrix would however lead to variables with large variances dominating the most important principal components. Also, changing the units of measurement (e.g., from ounces to pounds, or from feet to inches) would change the PCA solution. For this reason, it is often preferred to base the PCA solution on the eigenvectors and eigenvalues of the correlation matrix rather than the covariance matrix. This is equivalent to initially standardizing all variables and then performing the PCA is based on a correlation matrix [14].

(c) Calculate the eigenvectors and eigenvalues: Using concepts originating from linear algebra enables determining principal components stemming from the covariance matrix. An eigenvalue is a scalar that is used to transform (stretch) an eigenvector. The relevant equation is as follows:

$$Av = \lambda v,$$

where A is the square covariance matrix, v is an eigenvector, λ is a scalar which is the eigenvalue associated with the eigenvector of A matrix. A solution of this equation would yield λ eigenvalue:

$$\det(A - \lambda \mathrm{I}) = 0,$$

where det is the determinant, A is the covariance matrix, $\lambda\mathrm{I}$ is a scalar multiplying an identity matrix.

(d) Choose k eigenvectors with the largest eigenvalues: Sort the eigenvalues corresponding to eigenvalues from highest to lowest. In case the goal is to decrease the dimension to two, take the first two eigenvectors which are corresponding to the first two highest eigenvalues.

(e) Remodel the data: The final step uses the information from the eigenvectors of the covariance matrix to reorient data from the original axes to the ones that are now represented by the principal components:

$$y = W^T X$$

where W^T is the transpose of the matrix W, X is the eigenvector matrix and y is the transformed data set. Assuming our set of variables in the original data is $x_1, x_2, \dots, x_p$ after transformation the first principal component will be $Z_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p$, where $a_{11}, a_{21}, \dots, a_{p1}$ are the loadings of the first principal component. As earlier illustrated the loadings are values of the eigenvector of the covariance matrix. Since the eigenvalues are variances of the principal components, we can speak of "the proportion of variance explained" by the first k components [11]: proportion of variance $\frac{\lambda_1+\lambda_2+\cdots+\lambda_K}{\lambda_1+\lambda_2+\cdots+\lambda_P}$, where λ_1 refers to the variation explained by the first component, and so on [14].

To be able to come up with these principal components according to [14];

i. We can retain the first m components sufficient to explain a specified percentage (70% 80% 90% of the total variance of the original variables).
ii. Keep components whose eigenvalues are at least $\sum\frac{\lambda_i}{p}$ which is the average eigenvalue and also the average sample variance of the original variables, where $\lambda_i\ is$ a constant λ multiplying the number of factors i and p is the total number of observations in the data set.
iii. Use a scree plot of the eigenvalues λ_i, where λ is a constant and i is the number of factors. It always displays a downward curve. The point where the slope of the curve is leveling off (the elbow) indicates the number of factors that should be generated by the analysis as, Figure 10.

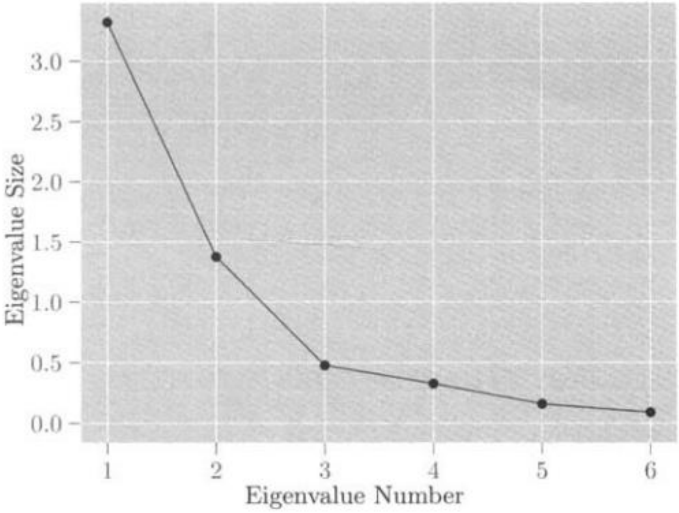


Figure 10: Scree Plot
Source: (Rencher and Christensen (2012)).

2. Clustering Analysis: It separates individual observations into groups based on the values for the p variables measured on each individual.

    a) Hierarchical Clustering

Agglomerative hierarchical clustering begins with n clusters, each containing a single object. At each stage, the two clusters that are "closest" are merged. As the stages iterate, there are n clusters, then n-1, and so on. By the last stage, there is 1 cluster containing all n objects, Figure 14.

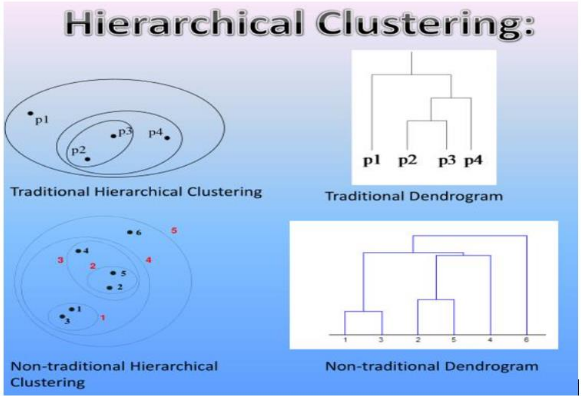

Hierarchical Clustering:
p1
p2
p3
p4
Traditional Hierarchical Clustering
p1 p2 p3 p4
Traditional Dendrogram
6
5
4
3
2
5
4
2
1
3
1
Non-traditional Hierarchical Clustering
1 3 2 5 4 6
Non-traditional Dendrogram

Figure 8: Hierarchical Clustering

Source: https://www.slideserve.com/lenci/partitional-clustering.

There are four common types of linkage: complete, average, single (Ward's), and centroid. A summary of these linkages is as follows [7 and 11];

- Complete: In this approach, all pairwise dissimilarities between the observations in the clusters are computed and the maximum one will be recorded [11].
- Single (Ward's): In this method, all pairwise dissimilarities between the clusters are computed and the minimum one will be recorded [11].
- Average: In this approach, all pairwise dissimilarities between the clusters are computed and the average of dissimilarities will be recorded.
- Centroid: In this technique, the dissimilarities between the mean vector of for cluster. A (centroid) and the mean vector for cluster B (centroid) are computed [11].

Steps in the agglomerative hierarchical clustering algorithm for grouping N objects according to [22];

- ➢ Start with N clusters, each containing a single entity and an N x N symmetric matrix of distances (or similarities) $D = d_{ik}$
- ➢ Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters U and V be d_{uv}
- ➢ Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by,
    - deleting the rows and columns corresponding to clusters U and V and
    - adding a row and column gives the distances between cluster (UV) and the remaining clusters.
- ➢ Repeat steps 2 and 3 a total of N −1 times. (All objects will be in a single cluster after the algorithm terminates.) Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

b) K-means

The k-means algorithm [10] begins by randomly allocating the n objects into k clusters (or randomly specifying k centroids). One at a time, the algorithm moves each object to the cluster whose centroid is closest to it, using the measure of closeness. When an object is moved, the centroids are immediately recalculated for the cluster gaining the object and the clutter losing it. The method repeatedly cycles through the objects until no reassignments of objects take place. The final clustering result will somewhat depend on the initial configuration
of the objects.
The k-means clustering results from a fundamental mathematical idea; Assume that $C_1, C_2, \dots, C_K$ represents sets including the observations clustered into K subgroups of the original data. These sets meet two properties [7 and 11];

a) $C_1 U\ C_2\ U, \dots, UC_{K'} = (1, \dots, n)$. It means the union of all clusters leads to the whole observation [11].

b) $C_k n\, C_k\ = \emptyset$ for all $k \neq k'$. It means clusters are pairwise and mutually exclusive [11].

The algorithm behind k-means clustering techniques [11] includes;

a) Randomly assign a number to each observation from 1 to K. This calls for an initial clustering of the observations [11].

b) Repeat the following process till the cluster assignments stop changing [11].

  i. For each of the K clusters, calculate the k^{th} cluster centroid which is the vector of the p feature means for the observations in the k^{th} cluster [11].

  ii. Use Euclidean distance for assigning each observation to the nearest Centroid [11].

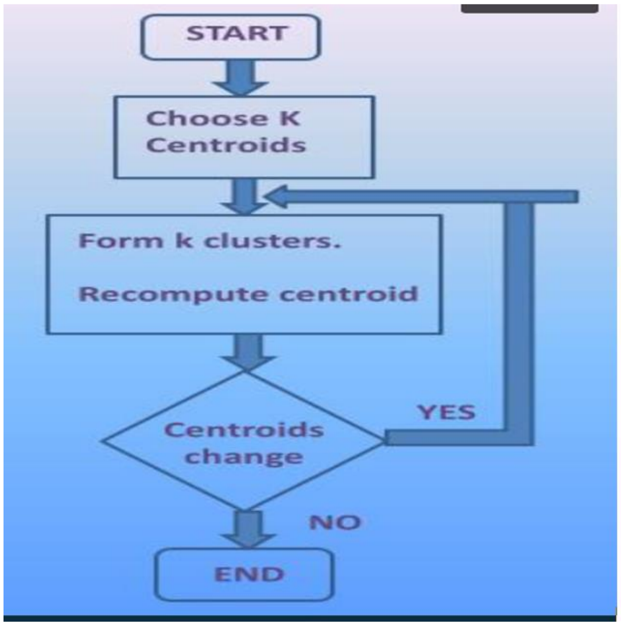


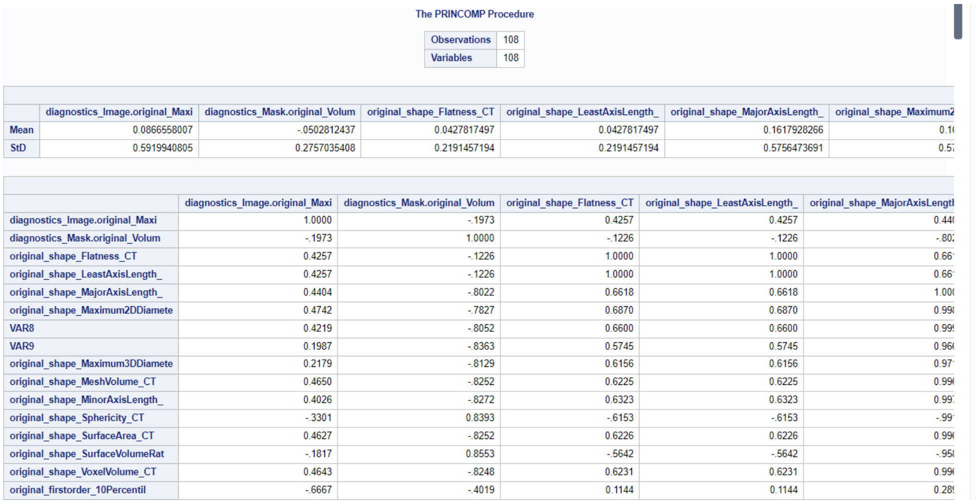
Figure 9: K-Means Clustering

Source: https://www.slideserve.com/lenci/partitional-clustering.

RESULTS FOR ANALYSES

a. Principal Component Analysis

With the data set on 108 features extracted from tumors of CT images of lung cancer patients, an illustration of how the reduction techniques were executed is shown. This particular analysis entailed both the use of SAS and partly R “softwares”. The data was standardized through SAS such that each variable had a mean of zero and a standard deviation of one. Principal components are computed from the correlation matrix, so the total variance is equal to the number of variables which is 108, Figure 10.



The PRINCOMP Procedure

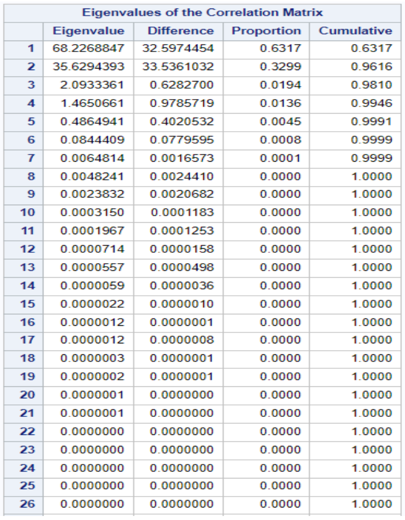
Observations	108
Variables	108

	diagnostics_Image.original_Maxi	diagnostics_Mask.original_Volum	original_shape_Flatness_CT	original_shape_LeastAxisLength_	original_shape_MajorAxisLength_	original_shape_Maximum2
Mean	0.0866558007	-.0502812437	0.0427817497	0.0427817497	0.1617928266	0.1
StD	0.5919940805	0.2757035408	0.2191457194	0.2191457194	0.5756473691	0.5

	diagnostics_Image.original_Maxi	diagnostics_Mask.original_Volum	original_shape_Flatness_CT	original_shape_LeastAxisLength_	original_shape_MajorAxisLengt
diagnostics_Image.original_Maxi	1.0000	-.1973	0.4257	0.4257	0.44
diagnostics_Mask.original_Volum	-.1973	1.0000	-.1226	-.1226	-.80
original_shape_Flatness_CT	0.4257	-.1226	1.0000	1.0000	0.66
original_shape_LeastAxisLength_	0.4257	-.1226	1.0000	1.0000	0.66
original_shape_MajorAxisLength_	0.4404	-.8022	0.6618	0.6618	1.00
original_shape_Maximum2DDiamete	0.4742	-.7827	0.6870	0.6870	0.99
VAR8	0.4219	-.8052	0.6600	0.6600	0.99
VAR9	0.1987	-.8363	0.5745	0.5745	0.96
original_shape_Maximum3DDiamete	0.2179	-.8129	0.6156	0.6156	0.97
original_shape_MeshVolume_CT	0.4650	-.8252	0.6225	0.6225	0.99
original_shape_MinorAxisLength_	0.4026	-.8272	0.6323	0.6323	0.99
original_shape_Sphericity_CT	-.3301	0.8393	-.6153	-.6153	-.99
original_shape_SurfaceArea_CT	0.4627	-.8252	0.6226	0.6226	0.99
original_shape_SurfaceVolumeRat	-.1817	0.8553	-.5642	-.5642	-.95
original_shape_VoxelVolume_CT	0.4643	-.8248	0.6231	0.6231	0.99
original_firstorder_10Percentil	-.6667	-.4019	0.1144	0.1144	0.28

Figure 10: Number of Observations and Simple Statistics

SAS software computes the principal components from the correlation matrix. By using eigenvalues as a way of selecting principal components, a summary table generated by the software informed the conclusion that the first, second and third principal components accounted for about 63.17%, 32.99% and 1.94% of the total variance respectively. Note that the sum of the eigenvalues is the total variance. The eigenvalues indicated that the first three components provide a good summary of the data accounting for 98.10% of the total variance while the rest of the components only account for less than 1.5% each, Figure 11.



Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	68.2268847	32.5974454	0.6317	0.6317
2	35.6294393	33.5361032	0.3299	0.9616
3	2.0933361	0.6282700	0.0194	0.9810
4	1.4650661	0.9785719	0.0136	0.9946
5	0.4864941	0.4020532	0.0045	0.9991
6	0.0844409	0.0779595	0.0008	0.9999
7	0.0064814	0.0016573	0.0001	0.9999
8	0.0048241	0.0024410	0.0000	1.0000
9	0.0023832	0.0020682	0.0000	1.0000
10	0.0003150	0.0001183	0.0000	1.0000
11	0.0001967	0.0001253	0.0000	1.0000
12	0.0000714	0.0000158	0.0000	1.0000
13	0.0000557	0.0000498	0.0000	1.0000
14	0.0000059	0.0000036	0.0000	1.0000
15	0.0000022	0.0000010	0.0000	1.0000
16	0.0000012	0.0000001	0.0000	1.0000
17	0.0000012	0.0000008	0.0000	1.0000
18	0.0000003	0.0000001	0.0000	1.0000
19	0.0000002	0.0000001	0.0000	1.0000
20	0.0000001	0.0000000	0.0000	1.0000
21	0.0000001	0.0000000	0.0000	1.0000
22	0.0000000	0.0000000	0.0000	1.0000
23	0.0000000	0.0000000	0.0000	1.0000
24	0.0000000	0.0000000	0.0000	1.0000
25	0.0000000	0.0000000	0.0000	1.0000
26	0.0000000	0.0000000	0.0000	1.0000

Figure 11: Principal Component Analysis of the first 26 features

A graphical representation of how many principal components should be retained to summarize our data was used. The graph is a scree plot of the eigenvalues λ_i against factor i. It always displays a downward curve. The point where the slope of the curve is clearly leveling off (the elbow) indicates the number of factors that should be generated by the analysis. The first three eigenvalues form a steep curve, followed by a bend and then a straight-line trend with a shallow slope [18]. The recommendation is to retain those eigenvalues in the steep curve before the first one on the straight line [11]. The scree plot confirmed an earlier conclusion made by the eigenvalues that 3 principal components were enough to explain variations from the original data which was about 98.10% in total. Additionally, the variance explained plot confirms that three components explain enough number of variations from the original data which is about 98.10% in total, Figure 12.

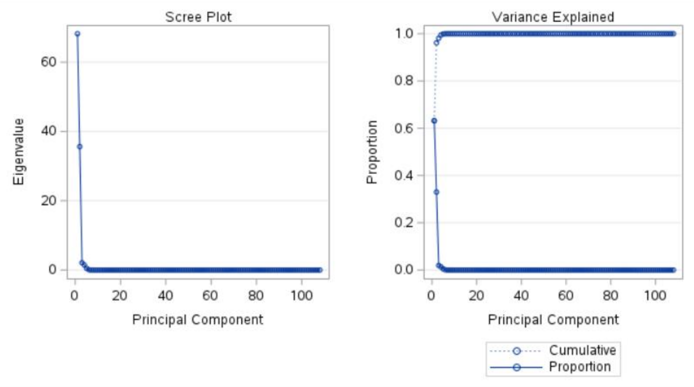


Figure 12: Scree Plot

To clearly specify which variables contributed most to the principal components, from the eigenvector's matrix, the first principal component Prin1 was written as a linear combination of the original variables, Figure 13.

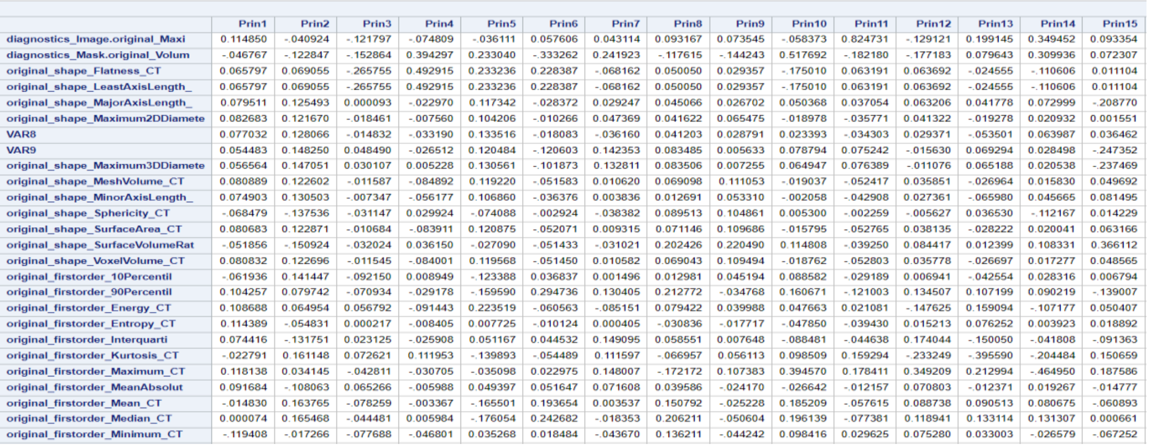
$$Prin_1 = 0.114850v_1 - 0.046767v_2 + \cdots - 0.014141v_{110}$$

The second principal component of Prin2 was,

$$Prin_2 = -0.040924v_1 - 0.122847v_2 + \cdots - 0.162658v_{110}$$

Finally, the third principal component of Prin3 was,

$$Prin_3 = -0.121797v_1 - 0.152864v_2 + \cdots + 0.098449v_{110}$$



	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15
diagnostics_Image.original_Maxi	0.114850	-.040924	-.121797	-.074809	-.036111	0.057606	0.043114	0.093167	0.073545	-.058373	0.824731	-.129121	0.199145	0.349452	0.093354
diagnostics_Mask.original_Volum	-.046767	-.122847	-.152864	0.394297	0.233040	-.333262	0.241923	-.117615	-.144243	0.517692	-.182180	-.177183	0.079643	0.309936	0.072307
original_shape_Flatness_CT	0.065797	0.069055	-.265755	0.492915	0.233236	0.228387	-.068162	0.050050	0.029357	-.175010	0.063191	0.063692	-.024555	-.110606	0.011104
original_shape_LeastAxisLength_	0.065797	0.069055	-.265755	0.492915	0.233236	0.228387	-.068162	0.050050	0.029357	-.175010	0.063191	0.063692	-.024555	-.110606	0.011104
original_shape_MajorAxisLength_	0.079511	0.125493	0.000093	-.022970	0.117342	-.028372	0.029247	0.045066	0.026702	0.050368	0.037054	0.063206	0.041778	0.072999	-.208770
original_shape_Maximum2DDiamete	0.082683	0.121670	-.018461	-.007560	0.104206	-.010266	0.047369	0.041622	0.065475	-.018978	-.035771	0.041322	-.019278	0.020932	0.001551
VAR8	0.077032	0.128066	-.014832	-.033190	0.133516	-.018083	-.036160	0.041203	0.028791	0.023393	-.034303	0.029371	-.053501	0.063987	0.036462
VAR9	0.054483	0.148250	0.048490	-.026512	0.120484	-.120603	0.142353	0.083485	0.005633	0.078794	0.075242	-.015630	0.069294	0.028498	-.247352
original_shape_Maximum3DDiamete	0.056564	0.147051	0.030107	0.005228	0.130561	-.101873	0.132811	0.083506	0.007255	0.064947	0.076389	-.011076	0.065188	0.020538	-.237469
original_shape_MeshVolume_CT	0.080889	0.122602	-.011587	-.084892	0.119220	-.051583	0.010620	0.069098	0.111053	-.019037	-.052417	0.035851	-.026964	0.015830	0.049692
original_shape_MinorAxisLength_	0.074903	0.130503	-.007347	-.056177	0.106860	-.036376	0.003836	0.012691	0.053310	-.002058	-.042908	0.027361	-.065980	0.045665	0.081495
original_shape_Sphericity_CT	-.068479	-.137536	-.031147	0.029924	-.074088	-.002924	-.038382	0.089513	0.104861	0.005300	-.002259	-.005627	0.036530	-.112167	0.014229
original_shape_SurfaceArea_CT	0.080683	0.122871	-.010684	-.083911	0.120875	-.052071	0.009315	0.071146	0.109686	-.015795	-.052765	0.038135	-.028222	0.020041	0.063166
original_shape_SurfaceVolumeRat	-.051856	-.150924	-.032024	0.036150	-.027090	-.051433	-.031021	0.202426	0.220490	0.114808	-.039250	0.084417	0.012399	0.108331	0.366112
original_shape_VoxelVolume_CT	0.080832	0.122696	-.011545	-.084001	0.119568	-.051450	0.010582	0.069043	0.109494	-.018762	-.052803	0.035778	-.026697	0.017277	0.048565
original_firstorder_10Percentil	-.061936	0.141447	-.092150	0.008949	-.123388	0.036837	0.001496	0.012981	0.045194	0.088582	-.029189	0.006941	-.042554	0.028316	0.006794
original_firstorder_90Percentil	0.104257	0.079742	-.070934	-.029178	-.159590	0.294736	0.130405	0.212772	-.034768	0.160671	-.121003	0.134507	0.107199	0.090219	-.139007
original_firstorder_Energy_CT	0.108688	0.064954	0.056792	-.091443	0.223519	-.060563	-.085151	0.079422	0.039988	0.047663	0.021081	-.147625	0.159094	-.107177	0.050407
original_firstorder_Entropy_CT	0.114389	-.054831	0.000217	-.008405	0.007725	-.010124	0.000405	-.030836	-.017717	-.047850	-.039430	0.015213	0.076252	0.003923	0.018892
original_firstorder_Interquarti	0.074416	-.131751	0.023125	-.025908	0.051167	0.044532	0.149095	0.058551	0.007648	-.088481	-.044638	0.174044	-.150050	-.041808	-.091363
original_firstorder_Kurtosis_CT	-.022791	0.161148	0.072621	0.111953	-.139893	-.054489	0.111597	-.066957	0.056113	0.098509	0.159294	-.233249	-.395590	-.204484	0.150659
original_firstorder_Maximum_CT	0.118138	0.034145	-.042811	-.030705	-.035098	0.022975	0.148007	-.172172	0.107383	0.394570	0.178411	0.349209	0.212994	-.464950	0.187586
original_firstorder_MeanAbsolut	0.091684	-.108063	0.065266	-.005988	0.049397	0.051647	0.071608	0.039586	-.024170	-.026642	-.012157	0.070803	-.012371	0.019267	-.014777
original_firstorder_Mean_CT	-.014830	0.163765	-.078259	-.003367	-.165501	0.193654	0.003537	0.150792	-.025228	0.185209	-.057615	0.088738	0.090513	0.080675	-.060893
original_firstorder_Median_CT	0.000074	0.165468	-.044481	0.005984	-.176054	0.242682	-.018353	0.206211	-.050604	0.196139	-.077381	0.118941	0.133114	0.131307	0.000661
original_firstorder_Minimum_CT	-.119408	-.017266	-.077688	-.046801	0.035268	0.018484	-.043670	0.136211	-.044242	0.098416	0.029625	0.075280	0.033003	-.026579	-.067252

Figure 13: Feature Loadings for 15 Principal Components

Also, we can illustrate the pairwise component score plots for the first components, with a 95% prediction ellipse overlaid on each scatter plot, Figures, 14,15 and 16. Figures 14 and 16 show the plot of the first components. The plots indicate regional trends in the plot of the first two components. Assuming components 2 and 3 are from a bivariate normal distribution, the ellipse identifies extracted features 39 which is the original_glcm_Contrast_CT as a possible outlier.

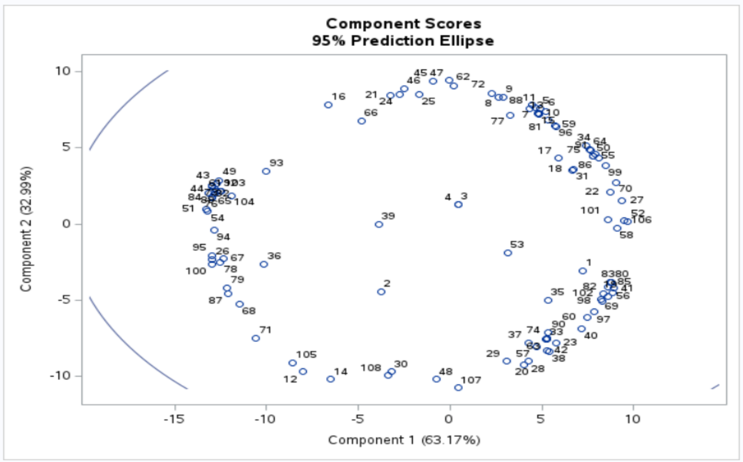


Figure 14:Plot of the First Two Component Scores

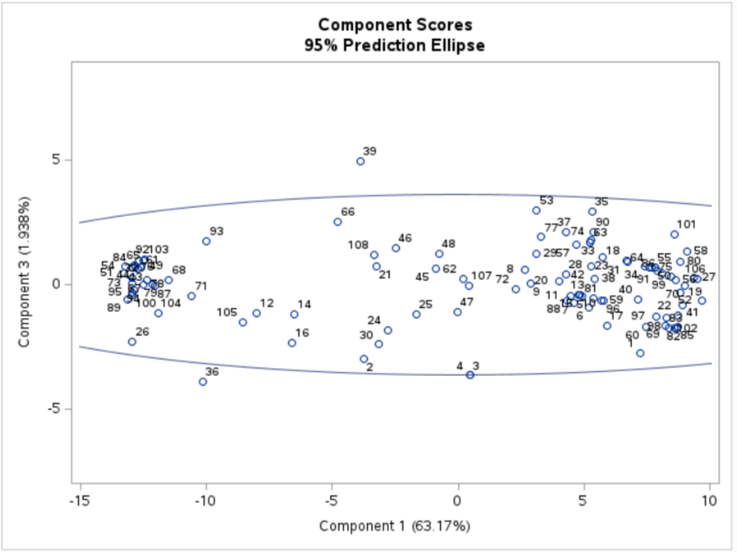


Figure 15:Plot of the First and Third Component Scores

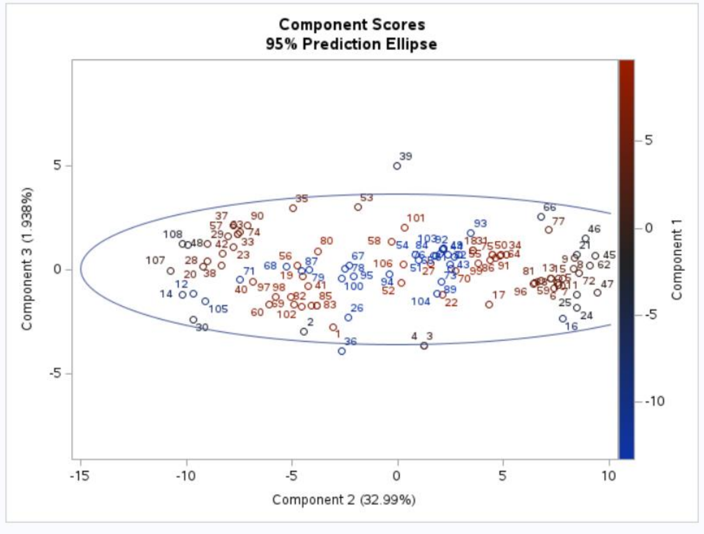
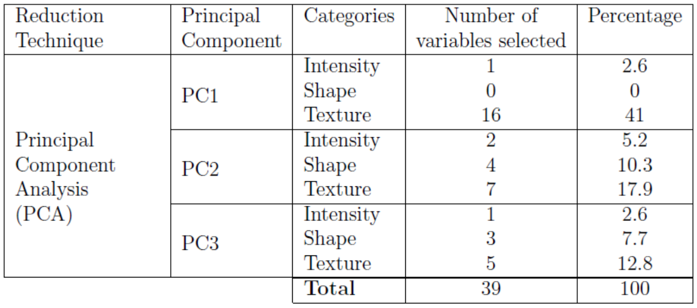


Figure 16: Plot of the Second and Third Component Scores

Since the main objective of the research was to end up with a smaller number of variables based on the principal components chosen, a low loadings filter was used based on the five-number summary for each variable in the principal components, a threshold value of 0.1 based on the maximum value of the loadings was chosen and features with loadings below 0.1 removed. This finally resulted in 39 features out of the 110 in the original data set with the texture-based being the most important. A summary of the selected features was, Table 1.



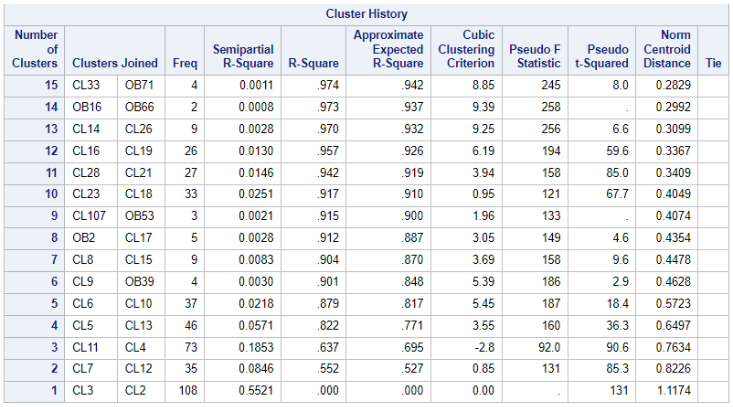
Reduction Technique	Principal Component	Categories	Number of variables selected	Percentage
Principal Component Analysis (PCA)	PC1	Intensity	1	2.6
		Shape	0	0
		Texture	16	41
	PC2	Intensity	2	5.2
		Shape	4	10.3
		Texture	7	17.9
	PC3	Intensity	1	2.6
		Shape	3	7.7
		Texture	5	12.8
		Total	39	100

Table 1: Features Selected through PCA

b. Clustering Analysis

i. Hierarchical Clustering

Examining the agglomerative hierarchical approach on the extracted features from lung cancer data by complete, average, and ward's minimum-variance clustering methods, SAS software was used. The results of cluster history are summarized in Figure 17. 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	Norm Centroid Distance	Tie
15	CL33	OB71	4	0.0011	.974	.942	8.85	245	8.0	0.2829	
14	OB16	OB66	2	0.0008	.973	.937	9.39	258	.	0.2992	
13	CL14	CL26	9	0.0028	.970	.932	9.25	256	6.6	0.3099	
12	CL16	CL19	26	0.0130	.957	.926	6.19	194	59.6	0.3367	
11	CL28	CL21	27	0.0146	.942	.919	3.94	158	85.0	0.3409	
10	CL23	CL18	33	0.0251	.917	.910	0.95	121	67.7	0.4049	
9	CL107	OB53	3	0.0021	.915	.900	1.96	133	.	0.4074	
8	OB2	CL17	5	0.0028	.912	.887	3.05	149	4.6	0.4354	
7	CL8	CL15	9	0.0083	.904	.870	3.69	158	9.6	0.4478	
6	CL9	OB39	4	0.0030	.901	.848	5.39	186	2.9	0.4628	
5	CL6	CL10	37	0.0218	.879	.817	5.45	187	18.4	0.5723	
4	CL5	CL13	46	0.0571	.822	.771	3.55	160	36.3	0.6497	
3	CL11	CL4	73	0.1853	.637	.695	-2.8	92.0	90.6	0.7634	
2	CL7	CL12	35	0.0846	.552	.527	0.85	131	85.3	0.8226	
1	CL3	CL2	108	0.5521	.000	.000	0.00	.	131	1.1174	

Figure 17: Cluster History

A figure summary, Figure 17 of the information includes the number and names of the clusters formed in the analysis. Each variable is identified either by a unique ID value or by CLn, where n corresponds to the cluster number. Figure 17 provides additional details, such as the count of observations in each new cluster and the semi-partial R square, which represents the reduction in variance resulting from merging two clusters. The R square, a measure of the proportion of variance explained by the clusters, is also displayed. For instance, when the data is divided into three clusters, the clusters account for approximately 63.7% of the variance. The ERSq column presents an approximate expected value of R square, which is calculated under the null hypothesis that the data exhibit a uniform distribution instead of distinct clusters. The next three columns display the values of the Cubic Clustering Criterion (CCC), Pseudo F (PSF), and *t*'' (PST2) statistics which assist in estimating the optimal number of clusters. The final column in Figure 17 indicates ties for minimum distance; a blank value implies no ties, whereas a tie suggests that the clusters could change by altering the order of observations

Interpretation of the CCC involves examining its values and patterns in relation to the number of clusters considered. The CCC assesses the fit of a clustering solution by considering the within-cluster sum of squares and the between-cluster sum of squares. It quantifies the trade-off between the compactness of clusters (minimizing within-cluster variation) and the separation between clusters (maximizing between-cluster variation). Since the CCC approach involves comparing the R-square obtained from a specific set of clusters with the R-square that would be obtained by clustering a uniformly distributed set of points. By examining the figure, it becomes apparent that there are two distinct maximum peaks observed at cluster number 5 and cluster number 13. Based on this observation, it is recommended to select the number of clusters between 5 and 13 which in this case is 5 clusters. To interpret the values of the pseudo t" 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. For this case, good clustering levels are observed at 3 clusters.

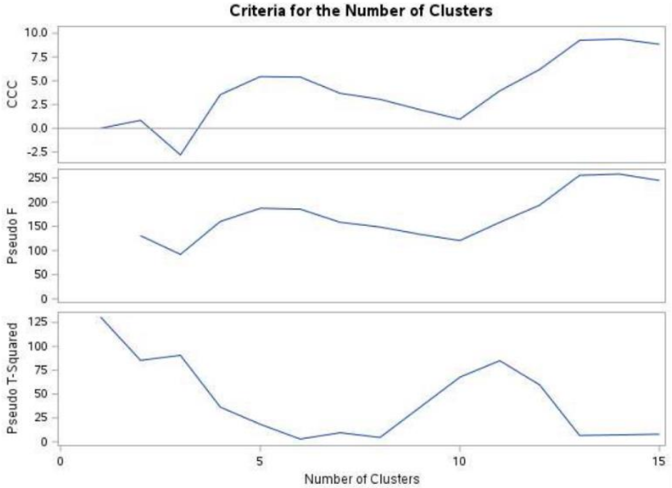


Figure 18: A plot of Statistics for Estimating the Number of Clusters

A dendrogram is a graphical representation of the hierarchy of clusters that shows the distance between clusters and the order in which they were merged. The height of the branches on the dendrogram represents the distance between the clusters. The closer the branches are to each other, the more similar the clusters are. As the number of branches grows to the left from the root, the R square approaches 1; the first three clusters (branches of the tree) account for over half of the variation (about 63.7%, from Figure 19). In other words, only three clusters are enough to explain over half of the variation.

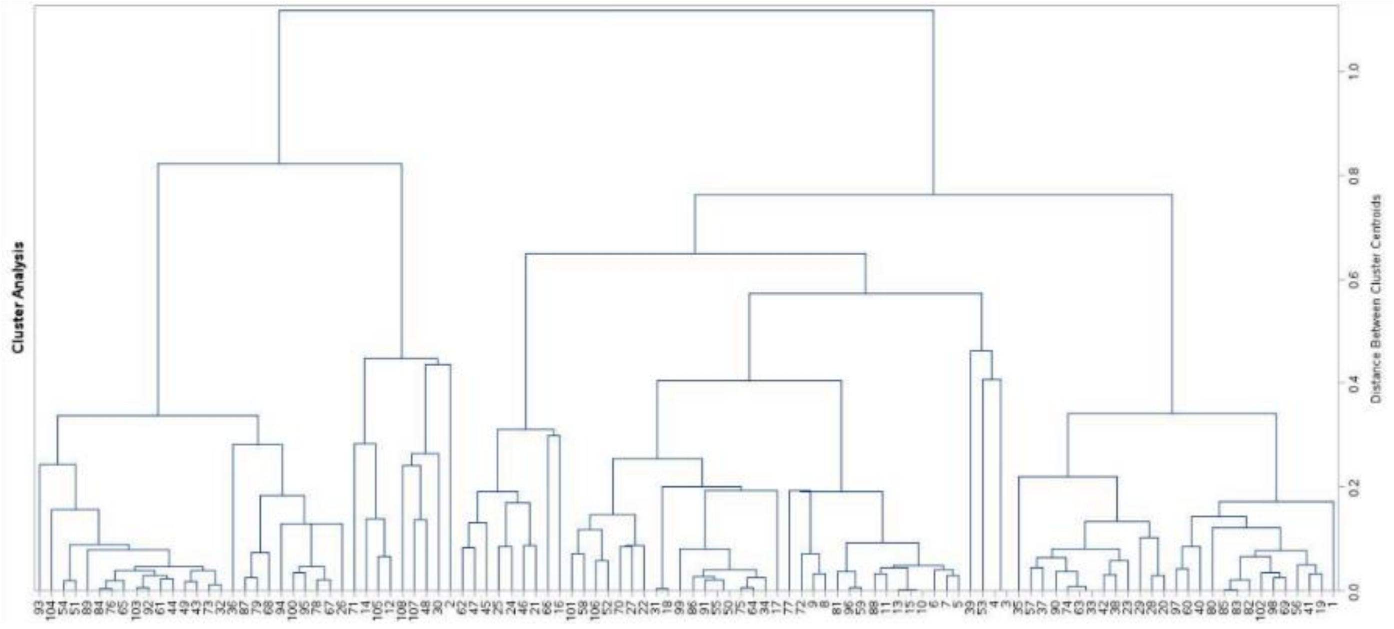


Figure 19: Dendrogram of Clusters versus R-Square Values

It was observed that cluster 1 started grouping with feature V_{93} and then V_{104} as it was trying to cluster features with the closest distance and the iteration continued till all the features were clustered. After determining the number of clusters, further analysis was undertaken to understand each cluster's characteristics. This was realized by identifying variables in each cluster and identifying any patterns or similarities. Through R software, the analysis progressed into using a variance filter on each cluster to identify features within each cluster with the most similarity. A five number summary on the selected feature matrix identified the 25th percentile averaging as 0.01 as an adequate threshold which was to drop every feature with a variance higher than 0.02. A summary of the selected features was as shown, Table 2. The most salient features observed are texture-based.

Reduction Technique	Clusters	Categories	Number of Variables selected	Percentage
Hierarchical Clustering	Clst1	Intensity Shape Texture	1 2 1	4.7 9.5 4.7
	Clst2	Intensity Shape Texture	2 1 0	9.5 4.7 0
	Clst3	Intensity Shape Texture	0 0 14	0 0 67
		Total	21	100

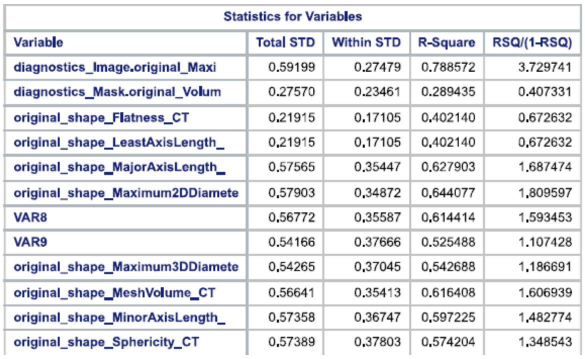
Table 2: Summary table of features selected through Hierarchical Clustering.

ii. K-means

For the k-means analysis, it is very important to find the optimal number of clusters beforehand before doing an analysis of the data set in this case by previous PCA and hierarchical clustering, the optimal number of clusters would be 3. When interpreting the output, it is important to consider the following:

- Cluster Assignments: Look at the cluster assignment of each observation, typically represented by a cluster ID or label. This indicates which cluster each observation belongs to. Analyze the distribution of observations across clusters to understand how they are grouped together.

- Cluster Profiles: Examine the cluster profiles or characteristics. This includes analyzing the means or centroids of the variables within each cluster. Look for variables with distinct values or large differences in means between clusters. These variables contribute the most to the separation of clusters and can help interpret the differences between them. The "within std" statistic helps understand the variability of the variables within each cluster. A higher "within std" value indicates that the data points within the cluster are more spread out or have greater variability for that particular variable. Conversely, a lower "within std" value suggests that the data points within the cluster are more tightly clustered or have less variability for that variable. Finally, RSQ/(1-RSQ) measure is often used as an indicator of how well a variable differentiates or separates the clusters. It can be thought of as a measure of the proportion of variation in the variable that is explained by the clustering. Higher values of "RSQ/(1-RSQ)" suggest that the variable has a stronger discriminatory power and is more effective in distinguishing the cluster

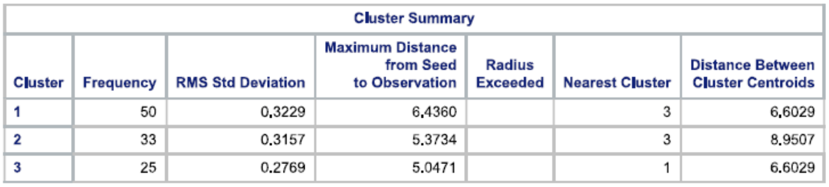


Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
diagnostics_Image.original_Maxi	0.59199	0.27479	0.788572	3.729741
diagnostics_Mask.original_Volum	0.27570	0.23461	0.289435	0.407331
original_shape_Flatness_CT	0.21915	0.17105	0.402140	0.672632
original_shape_LeastAxisLength_	0.21915	0.17105	0.402140	0.672632
original_shape_MajorAxisLength_	0.57565	0.35447	0.627903	1.687474
original_shape_Maximum2DDiamete	0.57903	0.34872	0.644077	1.809597
VAR8	0.56772	0.35587	0.614414	1.593453
VAR9	0.54166	0.37666	0.525488	1.107428
original_shape_Maximum3DDiamete	0.54265	0.37045	0.542688	1.186691
original_shape_MeshVolume_CT	0.56641	0.35413	0.616408	1.606939
original_shape_MinorAxisLength_	0.57358	0.36747	0.597225	1.482774
original_shape_Sphericity_CT	0.57389	0.37803	0.574204	1.348543

Figure 20: Statistics for Variables.

- Cluster Sizes: Assess the sizes of the clusters to understand their relative representation in the data set. Larger clusters may indicate dominant groups, while smaller clusters might represent more specific or unique patterns. Consider whether the cluster sizes align with your expectations or if there are imbalances that might require further investigation. SAS grouped 50 features in Cluster 1, 33 features in Cluster 2, and 25 features in Cluster 3. The table summary has a column for the Root Mean Square Standardized Distance (RMSSTD) whereby a lower RMSSTD value indicates that the observations within a cluster are more similar to each other, suggesting a tighter and more cohesive cluster such as cluster 3. Conversely, a higher RMSSTD value implies that the observations within a cluster are more dissimilar or scattered, indicating a less compact cluster seen in cluster 1. The same interpretation applies for the other columns for Maximum Distance from seed to observation and lastly for Distance between Cluster Centroids.

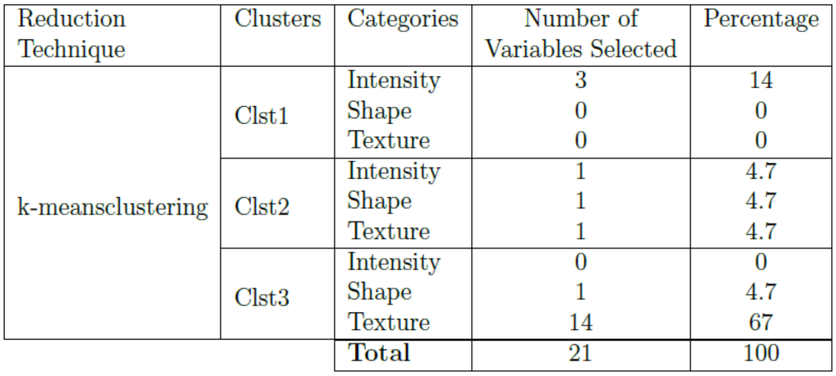


Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	50	0.3229	6.4360		3	6.6029
2	33	0.3157	5.3734		3	8.9507
3	25	0.2769	5.0471		1	6.6029

Figure 21: Cluster Summary.

The most salient features observed were texture-based. A summary table of features selected under k-means clustering, Table 3.



Reduction Technique	Clusters	Categories	Number of Variables Selected	Percentage
k-meansclustering	Clst1	Intensity	3	14
		Shape	0	0
		Texture	0	0
	Clst2	Intensity	1	4.7
		Shape	1	4.7
		Texture	1	4.7
	Clst3	Intensity	0	0
		Shape	1	4.7
		Texture	14	67
		Total	21	100

Table 3: Selected Features in k-means clustering.

CONCLUSION

The conclusion of any research is to ascertain if the research questions asked at the beginning of the research were successfully answered. This particular research was to find out if the number of variables could be reduced from 110 to a lesser number. The conclusion that the first, second and third principal components accounted for about 63.17%, 32.99% and 1.94% of the total variance respectively. The three components provided a good summary of the data accounting for 98.10% of the total variance. This finally resulted in 39 features out of the 110 in the original data set. A summary of the selected features was as follows; principal component one had a total of 17 features whereby 1 was the intensity and 16 were texture-based features, principal component two had a total of 13 features whereby 2 were shape, 4 intensity, and 7 texture-based features, and principal component three had a total of 9 features whereby 1 was the shape, 3 intensity, and 5 texture-based features. For clustering analysis, the agglomerative hierarchical clustering algorithm clustered the features into 3 clusters, 21 features were selected whereby 3 were intensity, 3 were shaped and 15 were texture-based features. K-means clustering algorithm with an initial cluster optimum cluster of 3, selected 21 features out of which 4 were intensity, 1 shape, and 15 texture-based features. Overall, all the analyses clearly outlined texture-based features as the most salient category of features.

REFERENCES

1. Velazquez ER Aerts HJ, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebers F, Rietbergen MM, Leemans C R, Dekker A, Quackenbush J, Gillies RJ, and Lambin P, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat Commun (2014), 5–4644.

2. Parnian Afshar, Arash Mohammadi, Konstantinos N Plataniotis, Anastasia Oikonomou, and Habib Benali, From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities, IEEE Signal Processing Magazine 36 (2019), no. 4, 132–160.

3. Michael Berry and Azlinah Mohamed, Supervised and unsupervised learning for data science, 01 2020.

4. B. Chen, L. Yang, R. Zhang, W. Luo, and W. Li, Radiomics: an overview in lung cancer management—a narrative review, Annals of Translational Medicine 8 (2020), 1191.

5. Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V. Miller, Steve Pieper, and Ron Kikinis, 3d slicer as an image computing platform for the quantitative imaging network, Magnetic Resonance Imaging 30 (2012), 1323–1341.

6. Yu-Ming Huang, Tsang-En Wang, Ming-Jen Chen, Ching-Chung Lin, Ching- Wei Chang, Hung-Chi Tai, Shih-Ming Hsu, and Yu-Jen Chen, Radiomics-based nomogram as predictive model for prognosis of

hepatocellular carcinoma with portal vein tumor thrombosis receiving radiotherapy, Frontiers in Oncology 12 (2022).

7. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An introduction to statistical learning, vol. 112, Springer, 2013.

8. Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aertsf, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies, Radiomics: the process and the challenges, Magn Reson Imag 30 (2012), 1234–48.

9. Zhou M., Scott J.and Chaudhury B.and Hall L., Goldgof D.and Yeom K. W.and Iv M.and Ou Y.and Kalpathy-Cramer, J. Napel, S. Gillies, R. Gevaert, O., and Gatenby R., Radiomics in brain tumor: Image assessment, quantitative feature descriptors, and machine-learning approaches and machine-learning approaches, American Journal of Neuroradiology 39 (2017), 208–216.

10. J MacQueen, Classification and analysis of multivariate observations, 5th Berkeley Symp. Math. Statist. Probability (1967), 281–297.

11. Zahed Mostafa and Skafyan Maryam, Application of feature selection and dimension reduction techniques on large-scale CT dataset for lung cancer diagnosis based on radiomics-sesug 2022 paper 222, (2022), 5–13.

12. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, and Aerts HJ, Radiomics: extracting more information from medical images using advanced feature analysis, Eur J Cancer 48 (2012), 441–6.

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

14. C Rencher Alvin, Methods of multivariate analysis, John Wiley&sons inc, Publication, Canada (2002).

15. Gillies RJ, Kinahan PE, and Hricak H, Radiomics: Images are more than pictures, they are data, Radiology 278 (2016), 563–77.

16. Jia Weikuan, Sun Meili, Lian Jian, and Hou Sujuan, Feature dimensionality reduction: a review, Complex & Intelligent Systems 8 (2022), 2198–6053.

17. Liu Z., Wang S., Dong D., Wei J., Fang C., Zhou X., Sun K., Li L., Li B., Wang M., and Tian J., The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges., Theranostics 9 (2019), 1303–1322.

18. Wang Z., Yang C., Han W., Sui X., Zheng F., Xue F., Xu X., Wu P., Chen Y., Gu W., W. Song, and Jiang J., Quantifying lung cancer heterogeneity using novel CT features a cross-institute study, Insights into Imaging 13 (2022).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please feel free to contact the authors at:

- Janet Akoth Kireta: KIRETA@etsu.edu
- Mostafa Zahed: zahedm@etsu.edu

APPENDIX

Abbreviations

For the sake of readability, the following is a list of the main abbreviations used in this paper:

CT	Computed Tomography
PET	PET Positron Emission Tomography
MRI	Magnetic Resonance Imaging
DICOM	Digital Imaging and Communications in Medicine
XML	Extensible Mark-up Language
OS	Overall Survival
PFS	Progression-Free Survival
NSCLS	Non-Small Cell Lung Cancer
PD-L1	Programmed Death L1igand 1
GLCM	Gray Level Co-occurrence Matrix
GLRLM	Gray Level Run Length Matrix
GLZLM	Gray Level Zone Length Matrix
NGTDM	Neighborhood Gray Tone Difference Matrix
MF	Minkowski Functional
LDA	Linear Discriminant Analysis
CCA	Canonical Correlation Analysis
NMF	Non-negative Matrix Factorization
FSA	Feature Selection Algorithm
FEA	Feature Extraction Algorithm

SAS CODE

```
FILENAME REFFILE '/home/u62120431/SAS JAN/corLung_Norm_mod.csv';

PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=COR;
        GETNAMES=YES;
RUN;

PROC CONTENTS DATA=COR; RUN;

PROC PRINCOMP DATA=COR;
RUN;

proc princomp DATA=COR plots= score(ellipse ncomp=3);
run;

/*Hierachical clustering*/
```

```
/* Perfoming Cluster Analysis */
ods graphics on;
proc cluster data=corrLung method = centroid ccc pseudo  print=15
outtree=Tree plots=den(height=rsq);
*var can1-can3;
*var diagnostics_Image.original_Maxim--original_ngtdm_Strength_CT;
run;

proc tree data=Tree out=New nclusters=3 noprint;
height_rsq_;
run;

ods graphics off;
/* Retaining 9 clusters */
proc tree data=Tree noprint ncl=6 out=out;
*copy diagnostics_Image.original_Maxim--original_ngtdm_Strength_CT;
run;
proc print data=out;
run;

/*k-means clustering*/
/* Run the  procedure */
proc fastclus data=corrLung out=output_data maxclusters=3; run;
```