

Unveiling the Diagnostic Potential of Radiomics-Based PET and CT Imaging through Statistical and Machine Learning Techniques

Mostafa Zahed; Janet Kireta; Victor Aderanti; East Tennessee State University,
Maryam Skafyan, University of Northern Colorado

ABSTRACT

Cancer remains a critical global health issue, with early detection and diagnosis essential for successful treatment. This research explores the integration of Radiomics techniques, utilizing advanced imaging modalities such as CT and PET-CT, to enhance the precision of medical image analysis for cancer detection and staging. Radiomics involves extracting quantitative features from medical images and applying machine learning for disease prediction, addressing the challenge of high-dimensional data through effective dimension reduction techniques. Utilizing datasets from medical universities in China, including a 2022 dataset from Harbin Medical University with CT and PET-CT images of 131 lung cancer patients, this study demonstrates the application of dimension reduction methods like PCA, K-means clustering, and Hierarchical Clustering using R, SAS, and Python. These techniques successfully reduced the number of features from 110 to 39 with PCA and to 21 with clustering. Texture-based features, particularly those in the second-order radiomics category, were identified as the most significant for lung cancer diagnosis and prediction. This research underscores the potential of Radiomics and unsupervised learning techniques to optimize feature extraction and reduction processes, enhancing the interpretability and efficiency of cancer prediction models from medical images. The findings highlight the crucial role of texture-based features in improving cancer diagnosis and treatment planning, providing valuable insights for future research and clinical applications.

INTRODUCTION

The study of cancer is vital due to its profound global impact and high mortality rates. At its core, cancer is characterized by the uncontrolled growth of abnormal cells, which can lead to tumor formation. Tumors can be benign or malignant, with malignant tumors having the potential to invade surrounding tissues and spread to other parts of the body (metastasis) [10]. In 2020, cancer was the second leading cause of death worldwide, resulting in nearly 10 million fatalities [37]. Early detection and diagnosis are crucial, and medical imaging plays a key role by enabling non-invasive visualization of the body's internal structures. Techniques such as CT, MRI, and PET scans are commonly used to identify and assess tumors across various types of cancer, including lung, breast, colorectal, and brain cancers, which were among the most common in 2020 [37, 10]. While research has traditionally focused on these prevalent types, it is essential to study all forms of cancer to develop more effective treatments and improve patient outcomes.

Radiomics, an emerging field, utilizes computational techniques to extract quantitative data from medical images. This approach provides deeper insights into tumor characteristics and aids in categorizing patients based on their likely response to treatment, paving the way for more personalized cancer therapies [37]. Radiomics involves the extraction of numerous quantitative features from medical images to define tumor phenotypes, transforming these images into high-dimensional data that can support clinical decision-making [3]. The foundation of radiomics is based on the idea that medical images contain valuable information about tumor morphology, physiology, and the surrounding environment, which, when properly extracted and analyzed, can enhance decision-making in oncology [2].

The radiomics process comprises four key steps: image acquisition, tumor segmentation, feature extraction, and data analysis [9]. Initially, standard clinical imaging modalities (such as CT, MRI, and PET) are utilized to obtain images following clinical protocols. The region of interest is then segmented—either manually, semi-automatically, or automatically—using specialized software. From these segmented images, hundreds of features related to the tumor's shape, intensity, texture, and wavelet patterns are extracted. These features are then statistically analyzed, often through machine learning techniques, to create predictive or prognostic models [10]. Radiomics has shown promise in improving predictions

related to treatment response, patient prognosis, and genetic profiling across various cancers, including lung, brain, breast, and prostate cancers [5].

Key applications of radiomics include tailoring therapies to individual patients, monitoring their response to treatment, and predicting clinical outcomes. However, there is still a need for validation of radiomic techniques, as well as the standardization of processes related to image acquisition, segmentation, and feature extraction [11]. Overall, radiomics offers a quantitative, non-invasive method to derive data from medical images, turning them into valuable resources for clinical decision-making and personalized cancer care [12]. Further research is essential to validate, standardize, and effectively implement these approaches.

The use of radiomics in cancer detection and analysis has seen significant growth in recent years, with multiple studies highlighting its potential benefits across various types of cancer. Radiomics, an emerging field in cancer research, involves the extraction and analysis of quantitative data from medical images. These features provide critical insights into tumor characteristics and are used in oncology for prediction, diagnosis, and treatment planning [4]. Previous research has shown that radiomic features can reveal tumor heterogeneity, which is crucial for understanding cancer progression and response to treatment. Several studies have developed and validated radiomic models for different cancers, such as lung, breast, colorectal, and prostate cancers. For instance, Aerts et al. (2020) demonstrated that radiomic features derived from CT images could effectively predict overall survival and treatment response in patients with non-small cell lung cancer, surpassing traditional clinical criteria [10].

Similarly, Parmar et al. (2015) explored the use of radiomics to predict distant metastases in breast cancer patients. The study found that radiomic features extracted from dynamic contrast-enhanced MRI could accurately forecast the development of distant metastases, providing critical information for treatment planning and personalized care [5]. Radiomics has also been used for predicting prognosis, assessing treatment response, and segmenting and classifying tumors. In another study, Bakr et al. (2018) applied radiomic features to segment glioblastoma tumors in MRI images, achieving higher accuracy than expert manual segmentation [6].

Moreover, radiomics has been integrated with other "omics" technologies, such as genomics and proteomics, to gain a more comprehensive understanding of cancer biology. For example, genomic mutations in cancer can affect the expression of various radiomic features, allowing for the identification of imaging signatures specific to different molecular subtypes. Dong et al. (2016) found that radiomic features derived from MRI could differentiate breast cancer subtypes based on genetic profiles, offering insights into tumor heterogeneity and potential treatment responses [7].

In this study, we will compare the results obtained from feature extraction and dimension reduction techniques for both CT and PET medical images.

RESEARCH METHODOLOGY

The study utilized a range of methods to explore the research question, including descriptive data analysis and various analytical techniques to examine and interpret the collected data, ultimately deriving meaningful insights.

DATA ANALYZE

The dataset was collected by Huiping Han, Funing Yang, and Rui Wang from Harbin Medical University in China [26]. It comprises CT and PET-CT DICOM images from 131 patients diagnosed with lung cancer. The radiomics workflow involves medical imaging, followed by segmentation to define the tumor region. From this segmented area, features related to tumor intensity, texture, and shape are extracted, which are then used for further analysis.

The XML annotation files, indicating the tumor locations, were provided by five experienced academic radiologists specializing in lung cancer. These annotations make the dataset a valuable resource for

developing algorithms for medical diagnosis. To extract the images and create bounding boxes around them, Python code was executed in the terminal. For feature extraction, Pyradiomics was used with the 3D Slicer medical software for 73 out of the 131 patients, as the provided annotation files were not compatible with all patients in Python. The dataset is publicly accessible on The Cancer Imaging Archive (TCIA) (<https://www.cancerimagingarchive.net/collection/lung-pet-ct-dx>).

To visualize the annotation boxes on the tumor within the DICOM images, Python code was executed in the terminal to extract the relevant images and enclose them within a bounding box [17].

There are various software tools available for feature extraction, including PyRadiomics, 3D Slicer, LIFEx, IBEX, QIFE, and RayPlus. Each of these tools comes with its own set of limitations and unique features, and there is no single standard for selecting the best one. The choice often depends on the researcher's approach and how they intend to handle medical images. For instance, 3D Slicer is an open-source software platform designed specifically for medical image informatics, image processing, and three-dimensional visualization. PyRadiomics is a Python package created to facilitate the extraction of radiomic features. LIFEx is a specialized software for extracting quantitative imaging features from various types of medical images. Ultimately, the selection of a software tool depends on the researcher's objectives and their level of expertise.

Extracted features are typically grouped into three main categories. The first category, known as first-order radiomics, includes features based on intensity and shape [16]. The second category, called second-order radiomics, involves texture-based features derived from several descriptive matrices, such as the Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Neighborhood Gray-Tone Difference Matrix (NGTDM), and Gray Level Zone Length Matrix (GLZLM).

The third category, higher-order radiomics, uses filters to extract features from images. This involves breaking down tumor images into various frequency domains (like horizontal, vertical, and diagonal) using wavelet techniques, which helps extract details about tumor shape, intensity, texture, and other characteristics. Fourier features are used to gather gradient information, while Minkowski Functionals (MF) are a common method for higher-order feature extraction, focusing on the arrangement of pixels that exceed a certain intensity threshold.

Out of 104 patients, features were successfully extracted from 98, as the annotation files for the remaining six patients were problematic. Segmentation was done using 3D Slicer, with the tumor highlighted in blue and pink, surrounded by a yellow circle.

The PyRadiomics package, integrated with 3D Slicer, was used to extract tumor features from the segmented images for all patients included in the study.

Following feature extraction, a 73 by 98 unsupervised data matrix was generated, where each row represented a patient, and each column corresponded to an extracted feature from the specified categories. The data matrix underwent min-max normalization to ensure that each feature was unitless and compatible with different feature distributions [17]. Due to the matrix's dimensions, conventional reduction techniques such as PCA, hierarchical clustering, and k-means struggled to handle this format. Consequently, the dataset was converted into a 73 by 68 square correlation matrix, enabling more effective application of dimension reduction methods.

In line with the research objectives, dimension reduction techniques were employed to enhance computational efficiency, avoid overfitting from an excessive number of features, and identify the most significant attributes for further analysis. Dimension reduction methods include feature extraction and feature selection. Feature extraction techniques are further classified into linear methods—Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Canonical Correlation Analysis (CCA)—and non-linear methods, such as Non-negative Matrix Factorization (NMF) and manifold-based methods. All of these aim to transform the original data into a new set of features that retain most of the information from the original dataset.

These methods involve both the extraction and selection of features [18]. Feature extraction techniques are classified into two primary categories: supervised and unsupervised learning. Supervised learning evaluates the relationship between features and their class labels, selecting features based on their

effectiveness in distinguishing between different classes. In contrast, unsupervised learning does not consider class labels and focuses on removing irrelevant features [19]. Given the unsupervised nature of the obtained data matrix, an extensive examination of unsupervised learning techniques was performed. This analysis covered both linear methods, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Canonical Correlation Analysis (CCA), as well as non-linear methods like Non-negative Matrix Factorization (NMF) and manifold-based techniques. The objective was to convert the original data into a new set of features that retains most of the original information. Similarly, feature selection methods aim to pinpoint the most crucial and informative features in the dataset. These methods include filter approaches, such as selecting based on low variance, information gain, chi-square tests, and Fisher's score; wrapper methods, which use techniques like forward selection, backward elimination, and exhaustive search; and embedded methods, such as regularization and assessing feature importance through random forest algorithms. The appropriate dimension reduction method depends on the dataset's specific characteristics and the research goals. Applying these methods helps researchers enhance computational efficiency, handle high-dimensional data challenges, and identify the most important features within the dataset, as illustrated in Figure 3.

Principal Component Analysis (PCA) is a widely used linear dimensionality reduction technique. PCA identifies principal components, which are orthogonal linear combinations of the original variables that capture the maximum variability in the data. The primary components are derived by solving the eigenvalue problem associated with the data's covariance matrix.

Non-linear dimensionality reduction is employed when it is assumed that the features do not lie in a linear space. In this study, manifold-based algorithms were selected. These algorithms project data into a lower-dimensional space, making it easier to visualize high-dimensional datasets, which is often challenging and less intuitive. Although random projection is the simplest method for dimensionality reduction, it may result in the loss of important structures within the data. The steps to achieve this reduction include:

1. Constructing a neighborhood graph using the data.
2. Calculating the shortest-path distances within this graph.
3. Applying multidimensional scaling to represent the data in fewer dimensions.

Hierarchical Clustering: Hierarchical clustering is a clustering technique that organizes objects into a hierarchical or tree-like structure based on their similarity. This method creates a dendrogram, where the leaves represent individual data points, and the branches represent the merging of similar clusters as we move up the hierarchy. Agglomerative hierarchical clustering begins with n clusters, each containing a single object. At each step, the two most similar clusters are merged, reducing the number of clusters from n to $n-1$, and so on, until only one cluster remains. Common linkage methods include single linkage, complete linkage, average linkage, and Ward's linkage.

K-means Clustering: K-means is a widely used unsupervised learning algorithm that partitions n observations into k clusters, where each observation is assigned to the cluster with the nearest mean. The algorithm starts by randomly assigning n objects to k clusters or selecting k initial centroids randomly. It iteratively reassigns each object to the closest cluster based on its centroid. After each reassignment, it updates the centroids for both the gaining and losing clusters. This process continues until there are no more changes in cluster assignments. The final clustering outcome can be influenced by the initial positioning of objects or centroids [30].

Feature Selection Algorithm

Feature selection is the process of identifying the most important features within a dataset to build models that are both accurate and efficient [25]. There are three main approaches to feature selection: filter, wrapper, and embedded methods. Filter methods assess the relevance of each feature by analyzing its statistical relationship with the target variable. Although there are various filter techniques available, our research focuses on variance and loadings [35]. These methods, especially when combined with other dimensionality reduction techniques, help in selecting features that show either minimal or substantial

variation across samples. Features with low variance are typically considered non-informative or redundant and may be excluded, while those with high variance can add uniqueness, especially in clustering analyses. However, removing high-variance features might simplify the analysis and improve accuracy, or it could complicate the data and lead to errors [35].

When using filters for dimensionality reduction, thresholds are set based on metrics such as the five-number summary, and features falling outside these thresholds are removed. Choosing these thresholds, however, requires careful consideration to balance the benefits of reducing dimensionality with the risk of losing important information.

RESULTS

This chapter seeks to thoroughly address these research questions by providing relevant data and analyses to enhance understanding. The original dataset contained 131 variables and 73 observations.

Using a dataset with 110 features extracted from CT images of lung cancer patients' tumors, the application of reduction techniques was demonstrated. This analysis was conducted using both SAS and R software in succession. The data was standardized in SAS so that each variable had a mean of zero and a standard deviation of one. A partial table of features before and after standardization is provided in Table 1.

Before Standardization			After Standardization	
Variable	Mean	Std Dev	Mean	Std Dev
diagnosticsImage.originalMaximumCT	2795.041	561.835	0.00	1.00
diagnosticsMask.originalVoxelN	1983.781	1983.918	0.00	1.00
diagnosticsMask.originalVolume	1.973	2.387	-0.00	1.00
originalshapeElongationCT	0.675	0.173	0.00	1.00
originalshapeFlatnessCT	0.003	0.0025	-0.00	1.00
originalshapeLeastAxisLengthC	0.363	3.119	-0.00	1.00
originalshapeMajorAxisLengthC	47.262	24.575	0.00	1.00
originalshapeMaximum2DDiameter	39.061	22.230	0.00	1.00
originalshapeMaximum2DDiameterRowCT	41.043	23.971	0.00	1.00
originalshapeMaximum2DDiameterSliceCT	53.315	27.032	-0.00	1.00
originalshapeMaximum3DDiameter	54.407	28.796	-0.00	1.00
originalshapeMeshVolumeCT	834.308	827.157	0.00	1.00
originalshapeMinorAxisLengthCT	31.080	16.408	0.00	1.00
originalshapeSphericityCT	0.213	0.078	0.00	1.00
originalshapeSurfaceAreaCT	2492.290	2419.104	-0.00	1.00
originalshapeSurfaceVolumeRatioCT	3.135	0.280	-0.00	1.00
originalshapeVoxelVolumeCT	857.009	836.936	0.00	1.00
originalfirstorder10PercentileCT	-286.601	211.378	0.00	1.00
originalfirstorder90PercentileCT	137.711	101.039	0.00	1.00
originalfirstorderEnergyCT	94411905.230	101889295.88	-0.00	1.00
originalfirstorderEntropyCT	4.505	0.547	-0.00	1.00
originalfirstorderInterquartileRangeCT	191.916	106.791	0.00	1.00

Table 1. A partial table of the means and standard deviations of some of the features before and after standardizing for CT extracted features.

Principal Components Analysis

Principal components are calculated from the correlation matrix, making the total variance equal to the number of variables, which is 110. SAS software was used to compute the principal components from the correlation matrix.

The remaining analysis was performed using R software to produce specific results, such as the scree plot. By using eigenvalues to select principal components, a summary table indicated that the first

principal component accounted for approximately 59.3% of the total variance, the second for about 29.0%, and the third for around 5.1%. It is important to note that the eigenvalues sum up to the total variance. The eigenvalues showed that three components provide a good summary of the data, accounting for 93% of the total variance, while the remaining components each account for less than 4%, as shown in Table 2.

Eigenvalue	Variance.percent	Cumulative.	variance.percent
Dim.1	12.11374	59.37771	59.37771
Dim.2	5.92293	29.03233	88.41004
Dim.3	1.037131	5.083689	93.49373
Dim.4	0.781042	3.828422	97.32215
Dim.5	0.249191	1.221458	98.5436
Dim.6	0.097194	0.476414	99.02002
Dim.7	0.076179	0.373403	99.39342
Dim.8	0.05263	0.257974	99.6514
Dim.9	0.034513	0.169171	99.82057
Dim.10	0.012128	0.059446	99.88001
Dim.11	0.007384	0.036196	99.91621
Dim.12	0.006432	0.031528	99.94774
Dim.13	0.004955	0.024286	99.97202
Dim.14	0.00153	0.0075	99.97952
Dim.15	0.000907	0.004445	99.98397
Dim.16	0.000797	0.003907	99.98788
Dim.17	0.000737	0.003613	99.99149
Dim.18	0.000444	0.002177	99.99367
Dim.19	0.000349	0.001713	99.99538
Dim.20	0.000254	0.001245	99.99662
Dim.21	0.000168	0.000822	99.99745
Dim.22	0.000143	0.000699	99.99815

Table 2. Principal components analysis of the first 22 features.

A graphical method was used to determine how many principal components should be retained to summarize the data. This graph, known as a scree plot, plots the eigenvalues (λ_i) against the corresponding factors (i) and typically shows a downward curve. The "elbow" point, where the slope of the curve levels off, indicates the number of factors to retain. The scree plot shows the first three eigenvalues forming a steep curve, followed by a bend and then a nearly straight line with a shallow slope [18]. The recommendation is to retain the eigenvalues in the steep part of the curve, before the first point on the straight line [18]. The scree plot confirmed the earlier conclusion based on eigenvalues that retaining 3 principal components was sufficient to account for approximately 93% of the variance in the original data, as shown in Figure 1.

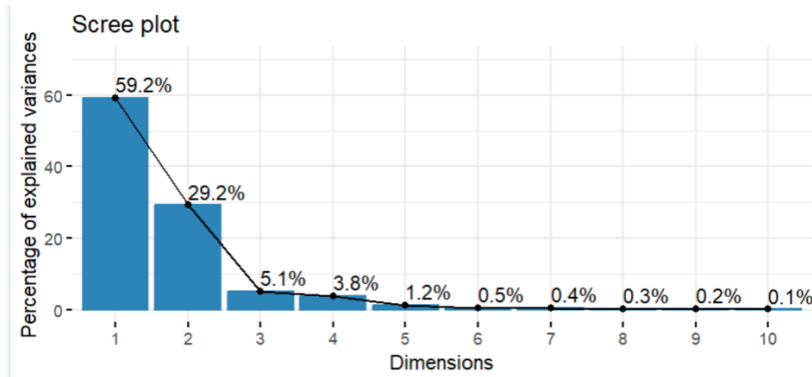


Figure 1. Scree Plot for CT images.

Given that the main goal of the research was to reduce the number of variables, a low loadings filter was applied based on the three chosen principal components. Using the five-number summary for each variable within these components, a threshold value of 0.1 was set, and features with loadings below this threshold were removed. This process reduced the original dataset from 110 features to 39. The selected features were distributed as follows: principal component one included 17 features—1 intensity-based and 16 texture-based; principal component two comprised 13 features—2 shape, 4 intensity, and 7 texture-based; and principal component three contained 9 features—1 shape, 3 intensity, and 5 texture-based. Based on these results, we concluded that the most important feature category according to PCA was the texture feature category. A summary of the selected PCA features is provided in Tables 3 and 4.

Principal Component 1	Principal Component 2	Principal Component 3
original_firstorder_Minimum_CT- intensity feature	original_shape_MinorAxisLength_CT- shape feature	original_shape_VoxelVolume_CT- shape feature
original_gldm_Correlation_CT- texture feature	original_shape_SurfaceArea_CT- shape feature	original_firstorder_MeanAbsoluteDe- viation_CT- intensity featuresn
original_gldm_DifferenceAverage_CT- texture feature	original_firstorder_Energy_CT- intensity feature	original_firstorder_Median_CT- intensity feature
original_gldm_JointEnergy_CT- texture feature	original_firstorder_Maximum_CT- intensity feature	original_firstorder_RootMeanSquare d_CT- intensity feature
original_gldm_MCC_CT- texture feature	original_firstorder_Range_CT- intensity feature	original_gldm_ClusterProminence_C T- texture feature
original_gldm_DependenceNonUnif ormityNormalized_CT - texture	original_firstorder_RobustMeanAbsol uteDeviation_CT - intensity feature	original_gldm_RunLengthNonUnifor mity_CT- texture feature
original_gldm_LowGrayLevelEmpha sis_CT- texture feature	original_gldm_ClusterTendency_CT- texture feature	original_gldm_RunLengthNonUnifor mityNormalized_CT- texture feature
original_gldm_SmallDependenceEm phasis_CT - texture feature	original_gldm_DifferenceEntropy_CT- texture feature	original_gldm_RunVariance_CT- texture feature
original_gldm_LowGrayLevelRunEm phasis_CT - texture feature	original_gldm_Imc1_CT - texture feature	original_gldm_ZoneEntropy_CT- texture feature
original_gldm_RunLengthNonUnifor mity_CT- texture feature	original_gldm_JointEntropy_CT- texture feature	
original_gldm_RunLengthNonUnifor mityNormalized_CT - texture feature	original_ngtdm_Busyness_CT - texture feature	
original_gldm_RunVariance_CT - texture feature	original_ngtdm_Complexity_CT - texture feature	
original_gldm_SizeZoneNonUniform ity_CT- texture feature	original_ngtdm_Contrast_CT- texture feature	
original_gldm_SizeZoneNonUniform ityNormalized_CT- texture feature		
original_gldm_SmallAreaEmphasis_ CT- texture feature		
original_gldm_ZoneEntropy_CT- texture feature		
original_ngtdm_Coarseness_CT- texture feature		

Table 3. A list of features selected for CT images in each principal component.

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 4. Summary of Selected Features for CT Images.

Similarly, for PET images, the dataset underwent Principal Component Analysis (PCA) to identify patterns and relationships among variables. A crucial part of PCA is understanding the variation explained by each principal component, which we illustrated using a scree plot. The scree plot, which shows the proportion of total variance explained by each component, indicated that the first five components accounted for approximately 45.1%, 14.1%, 12.1%, 6.7%, and 4.7% of the variance, respectively. The optimal number of principal components to retain depends heavily on the location of the "elbow" in the scree plot, which marks a notable decline in the amount of variance explained by subsequent components. Based on these percentages, the first five components are highly informative and capture a substantial portion of the data's variability, as shown in Figure 2.

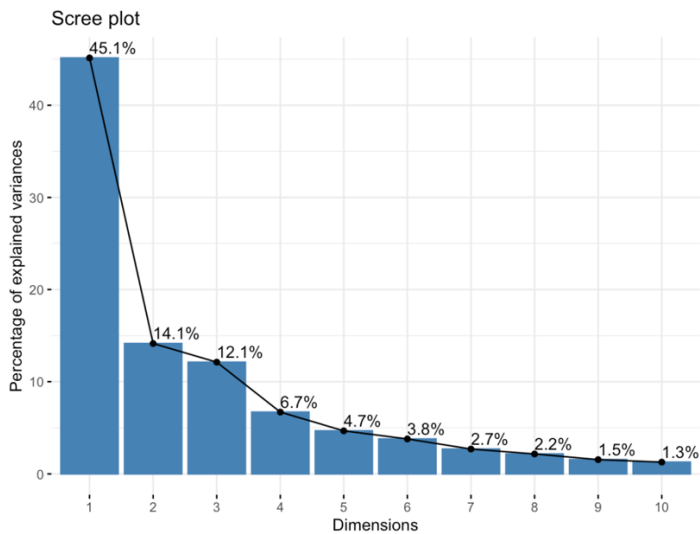


Figure 2. Scree Plot for PET images.

Since the main goal of the research was to reduce the number of variables, a low loadings filter was applied using the five selected principal components. This process involved setting a threshold value of 0.15, based on the five-number summary for each variable within these principal components, and removing features with loadings below this threshold, as shown in Table 5.

Reduction Technique	Principal Component	Categories	Number of Variable Selected	Percentage
Principal Component Analysis (PCA)	PC1	Texture	10	38.5
		Shape	1	3.8
		Intensity	3	11.5
	PC2	Texture	1	3.8
		Shape	0	0
		Intensity	0	0
	PC3	Texture	0	0
		Shape	2	7.7
		Intensity	0	0
	PC4	Texture	5	19.2
		Shape	0	0
		Intensity	0	0
	PC5	Texture	0	0
		Shape	4	15.4
		Intensity	0	0
			26	100

Table 5. Features selected through PCA for PET Images.

Hierarchical Clustering Approach

To analyze the extracted features from lung cancer data using the agglomerative hierarchical approach, complete, average, and Ward's minimum-variance clustering methods were applied using R software. A dendrogram, which is a graphical representation of the cluster hierarchy, illustrates the distance between clusters and the sequence in which they were merged. The height of the branches on the dendrogram represents the distance between clusters—the closer the branches, the more similar the clusters are, as shown in Figures 3.

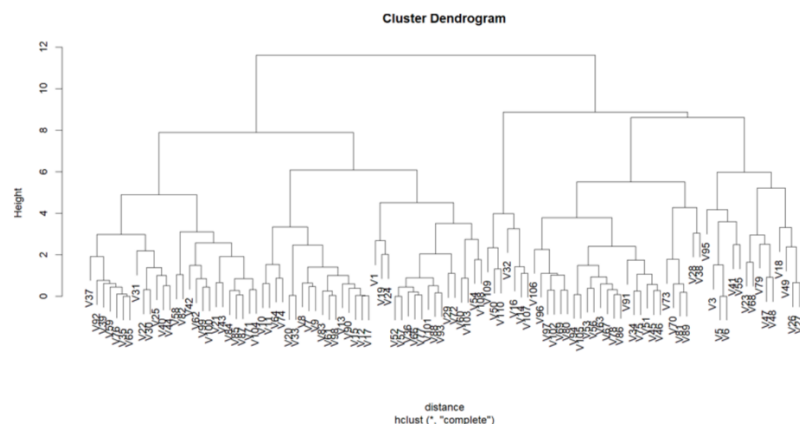


Figure 3. Cluster dendrogram of all features for CT.

The hierarchical clustering algorithm grouped the features into 3 clusters, Table 6.

Features	Cluster 1	Cluster 2	Cluster 3
V1	1	0	0
V10	0	1	0
V100	1	0	0
V101	0	1	0
V102	0	0	1
V103	0	1	0
V104	1	0	0
V105	0	0	1
V106	0	0	1
V107	0	0	1
V108	0	1	0
V109	1	0	0
V11	0	1	0
V110	1	0	0
V12	0	1	0
V13	0	1	0
V14	0	0	1
V15	0	1	0
V16	0	0	1
V17	0	1	0
V19	0	1	0
V20	0	1	0
V21	1	0	0
V23	0	1	0
V24	0	1	0
V25	1	0	0
V26	0	1	0
V27	0	1	0
V28	0	0	1
V29	0	1	0

Table 6. Features per cluster HC, CT images.

The feature names were labeled as V1, V2, ..., V110 instead of their original variable names. It was observed that cluster 2 began grouping with features V10 and then V101, attempting to cluster features with the closest distance, and this process continued until all features were clustered. After determining the number of clusters, the three clusters were further analyzed to understand their characteristics by identifying the variables within each cluster and recognizing any patterns or similarities. Using the `cutree()` function in R, 48 features were grouped into cluster 1, 33 features into cluster 2, and 27 features into cluster 3.

To assess if the chosen number of clusters was appropriate, R was used to generate the following:

a) A silhouette plot, which is a graph used to interpret the results of clustering algorithms, including hierarchical clustering. In this plot, each data point is represented by a vertical line or bar that reflects its silhouette coefficient—a measure of how similar the point is to its own cluster compared to other clusters. The silhouette coefficient ranges from -1 to 1, with higher values indicating that a point is well-matched to its own cluster and poorly matched to others. The silhouette plot typically displays each data point sorted by its cluster assignment, with the silhouette coefficient plotted against the index of the data point, as shown in Figures 4, 5, and 6.

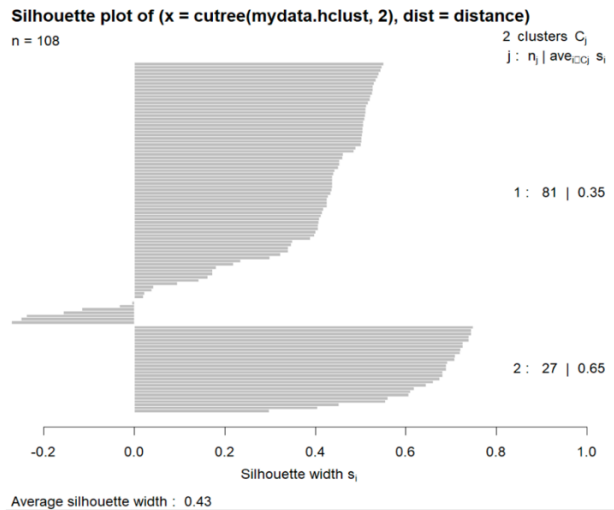


Figure 4. Silhouette plot 2 clusters for CT.

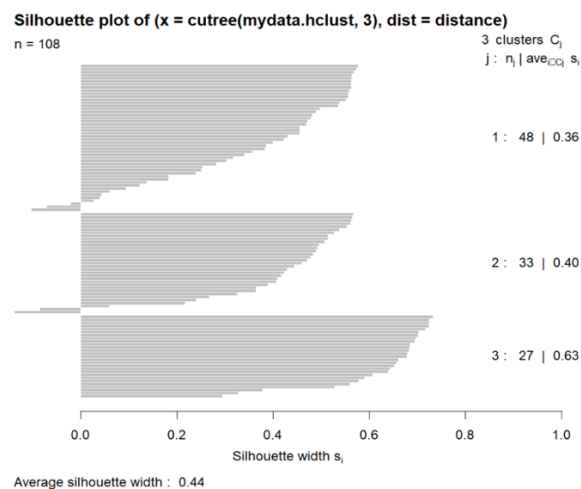


Figure 5. Silhouette plot 3 clusters for CT.

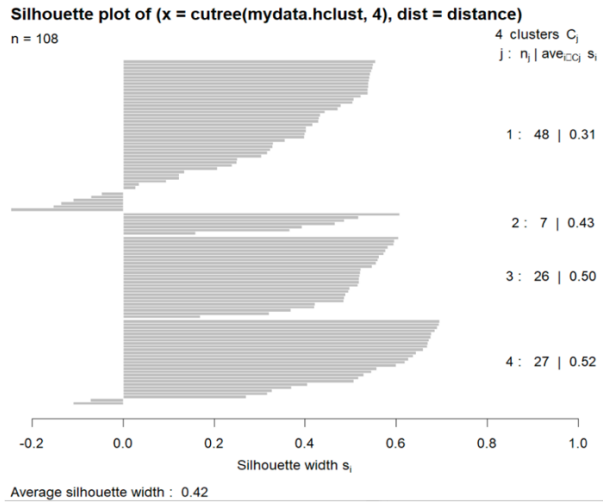


Figure 6. Silhouette plot 4 clusters for CT.

Interpreting silhouette plots involves the following considerations:

- **Looking for well-defined clusters:** If there are clear gaps between the bars for different clusters, it suggests that the clustering algorithm has successfully separated the data into distinct groups. In our silhouette plots, Figures 25 and 26 show clear gaps, indicating successful clustering, while Figure 27 lacks a clear gap, suggesting that the clustering algorithm did not effectively separate the clusters.
- **Looking for overlap between clusters:** Significant overlap between the bars for different clusters indicates that the clustering algorithm may not have effectively separated the data into distinct groups. Figure 27 shows overlap, whereas Figures 25 and 26 have no overlaps, making these two better cluster configurations.
- **Assessing the overall quality of the clustering:** The average silhouette coefficient for all data points serves as a measure of the overall quality of the clustering. A high average silhouette coefficient indicates that data points are well-matched to their respective clusters, whereas a low average silhouette coefficient suggests that the clustering algorithm did not effectively separate the data into distinct groups. Among the clustering options, the plot with 3 clusters has an average silhouette coefficient of 4.3, indicating that it is a more suitable number of clusters.

a) The **Elbow Method** examines the total within-cluster sum of squares (WSS) as a function of the number of clusters. Similar to the scree plot in PCA, the "knee" point in the plot is considered an indicator of the appropriate number of clusters. Based on this plot, it was determined that 3 clusters would be sufficient, as shown in Figure 7.

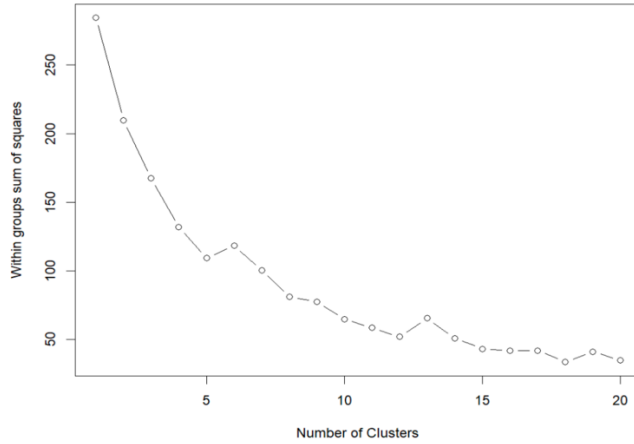


Figure 7. Elbow method for optimal number of clusters for CT.

similar features within them. A five-number summary of the selected feature matrix determined that the 25th percentile, with an average of 0.01, was an appropriate threshold, leading to the removal of features with variance exceeding 0.02. The final selection included 21 features: 4 in Clst1, 3 in Clst2, and 14 in Clst3. In Cluster 1, there was 1 intensity feature, 2 shape features, and 1 texture feature; Cluster 2 contained 2 intensity and 1 shape feature, with no texture features; and Cluster 3 comprised 14 texture features, with no shape or intensity features. The findings showed that texture features were the most dominant, making up the largest portion of the new data matrix. These results are summarized in Table 7.

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

Table 7. Summary table of features selected through hierarchical clustering, CT images.

For PET images, an analysis was performed using the agglomerative hierarchical technique on characteristics derived from lung cancer data, applying complete, average, and Ward's minimum-variance clustering methods in the R programming language. A dendrogram, which visually represents the cluster hierarchy, shows the distances between clusters and the order in which they merge. The height of the branches on the dendrogram indicates the distance between clusters, while the closeness of the branches reflects the similarity between them. Visual representations can be found in Figures 8 and 9.

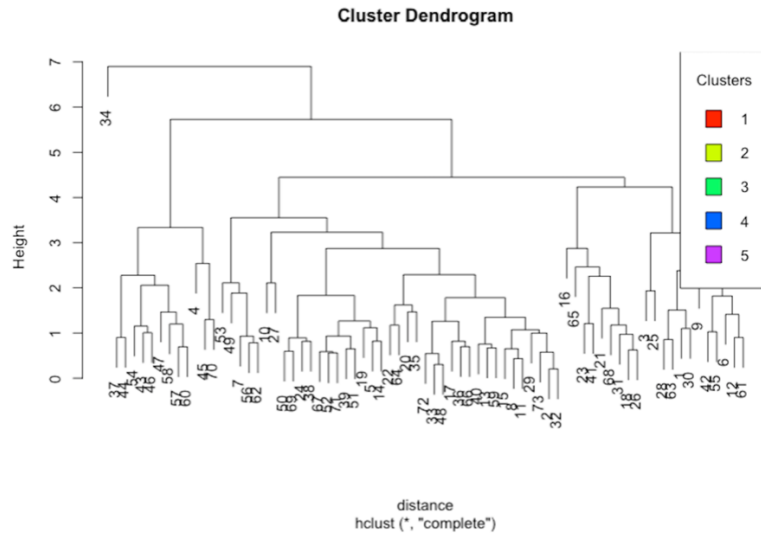


Figure 8. Cluster dendrogram of all features for PET.

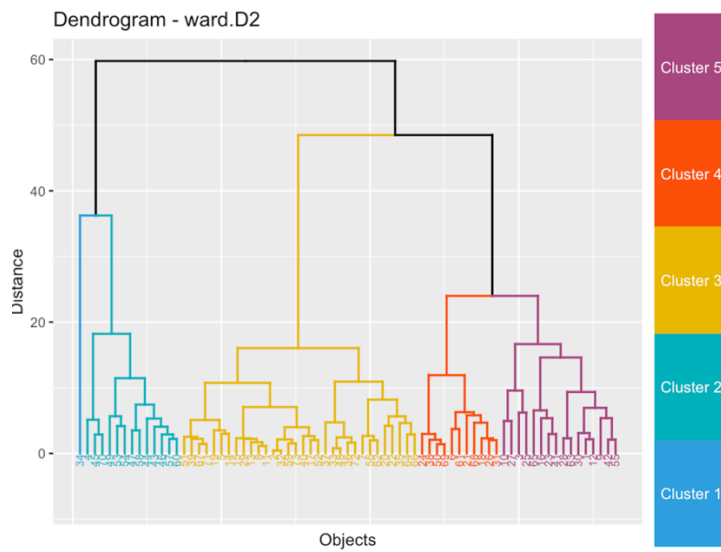


Figure 9. Dendrogram illustrating the clustering of all features.

The analysis advanced by applying a variance filter to each cluster to identify the features with the most similarity. A five-number summary of the selected feature matrix revealed that the 25th percentile, averaging 0.01, served as an appropriate threshold. Features with variance greater than 0.05 were excluded, as shown in Table 8.

Summary Statistics for all selections							Variance V
	Min	Q_1	Median	Mean	Q_3	Max	$V \geq 0.05$
Cluster 1	0.02881	0.04005	0.04957	0.05149	0.06141	0.07973	5
Cluster 2	0.01358	0.02316	0.02708	0.03455	0.04856	0.06800	6
Cluster 3	0.02476	0.04068	0.04925	0.04868	0.05726	0.07144	2
Cluster 4	0.01706	0.04446	0.05098	0.04816	0.05476	0.07063	4

Table 8. Summary Statistics and Loadings for all Selections, PET images.

The selected features included 17 in total: 5 in Cluster 1, 6 in Cluster 2, 2 in Cluster 3, and 4 in Cluster 4. Cluster 1 had 2 intensity, 1 shape, and 2 texture features; Cluster 2 included 1 intensity, 1 shape, and 4 texture features; Cluster 3 contained 2 texture features with no shape or intensity; and Cluster 4 had 1 texture and 3 intensity features, with no shape features. The findings indicated that texture-based features were the most dominant, contributing the largest percentage to the new data matrix, as summarized in Table 9.

Reduction Technique	Principal Component	Categories	Number of Variable Selected	Percentage
Hierarchical Clustering	Cluster 1	Texture	2	11.8
		Shape	1	5.9
		Intensity	2	11.8
	Cluster 2	Texture	4	23.5
		Shape	1	5.9
		Intensity	1	5.9
	Cluster 3	Texture	2	11.8
		Shape	0	0
		Intensity	0	0
	Cluster 4	Texture	1	5.9
		Shape	0	0
		Intensity	3	17.6
	Total	17	100	

Table 9. Summary table of selected features through hierarchical clustering, PET images.

K-Means

For the k-means analysis, determining the optimal number of clusters is crucial before analyzing the dataset. The analysis was conducted using R software, starting with loading the libraries factextra and cluster. The cancer data, previously loaded and named df, was cleaned by removing rows with missing values using na.omit(df). Each variable was scaled to have a mean of 0 and a standard deviation of 1 using the scale(df) function. To find the optimal number of clusters, two plots were used:

- Number of Clusters vs. Total Within Sum of Squares:** The fviz-nbclust() function created a plot showing the number of clusters versus the total within sum of squares. The plot, shown in Figure 25, indicates an "elbow" at $k = 3$, where the sum of squares levels off, suggesting this as the optimal number of clusters (Figure 10).

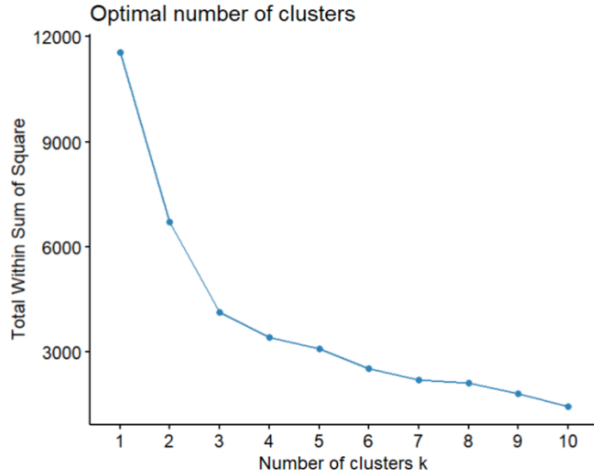


Figure 10. A plot of the total within sum of squares vs. the number of clusters.

b) **Gap Statistic Method:** This method compares intra-cluster variation for different values of k with expected values from a distribution without clustering. The gap statistic for each number of clusters was calculated using the `clusGap()` function, with a plot generated using `fviz-gap-stat()`. The plot showed the highest gap statistic at $k = 3$ clusters (Figure 11).

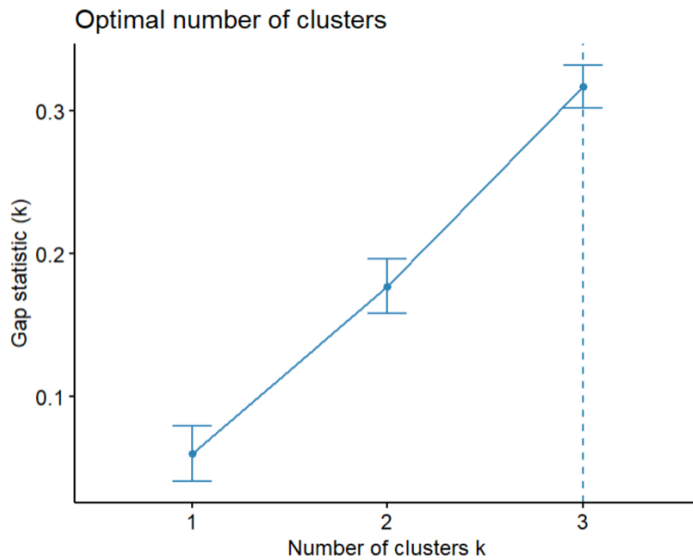


Figure 11. A plot of gap statistic versus clusters, CT images.

A variance filter was applied to each cluster in the k-means clustering process. Features with variance less than 0.02 were removed, resulting in 21 selected variables: 3 in Cluster 1, 3 in Cluster 2, and 15 in Cluster 3. Cluster 1 consisted of 3 intensity-based features, with no shape or texture features. Cluster 2 had 1 intensity, 1 shape, and 1 texture feature, while Cluster 3 included 1 shape and 14 texture features, with no intensity features. The analysis concluded that texture-based features were the most significant, as summarized in Table 10.

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 10. Summary table of features selected through k-means clustering, CT images.

For PET images, identifying the optimal number of clusters was essential before performing k-means analysis. The analysis was conducted in R, starting with the loading of factoextra and cluster libraries. The cancer dataset, named df, was cleaned of missing values using the na.omit(df) function, a common data management practice in data science.

For PET analysis, determining the optimal number of clusters is essential before applying k-means clustering to a dataset. The analysis was performed in R, beginning with the loading of the factoextra and cluster libraries. The cancer dataset, referred to as df, was cleaned of missing values using the na.omit(df) function, demonstrating a common data management practice in data science.

To identify the best number of clusters, the gap statistic method was used. This metric assesses intra-cluster variation for different values of k and compares them to expected values from a distribution without clustering. The gap statistic was calculated for each number of clusters using the clusGap() function from the cluster package, and a plot of clusters versus gap statistic was generated with the fviz-gap-stat() function. The analysis revealed that the gap statistic peaked at 4 clusters, as shown in Figure 12.

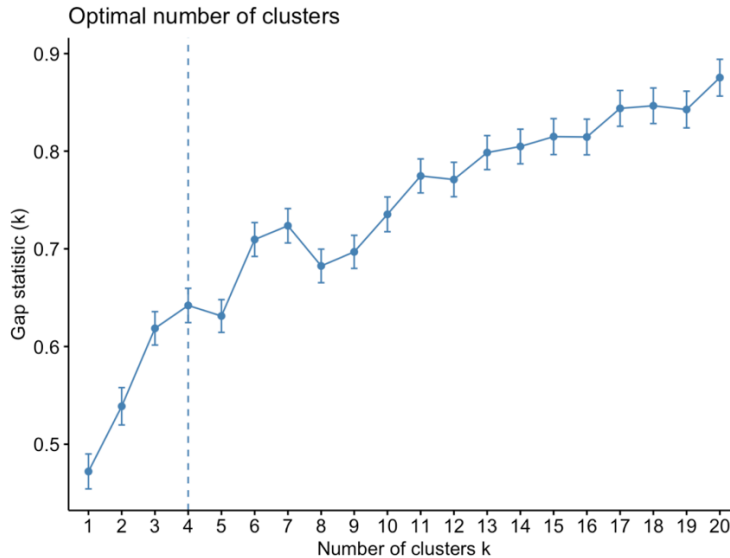


Figure 12. A plot of gap statistic versus clusters, PET images.

CONCLUSION

For CT analysis, the primary goal of the research was to determine if the number of variables could be reduced from 110 to a smaller set. After the analysis, 39 features were selected for PCA, and 21 features for both hierarchical and k-means clustering. Another key research question was to identify the most significant feature category. All analyses confirmed that texture-based features were the most important. The research successfully met its objective by reducing the feature set to between 20 and 40, reducing dimensionality by over 75%. This resulted in a data matrix of 74 by 39 for PCA and 74 by 21 for clustering analyses.

Specifically, for PCA, eigenvalues showed that three principal components summarized 93% of the total variance, with the remaining components contributing less than 4% each. A loadings filter with a threshold of 0.1, based on a five-number summary, selected 39 features. Principal Component 1 included 17 features (1 intensity, 16 texture), Principal Component 2 had 13 features (2 shape, 4 intensity, 7 texture), and Principal Component 3 had 9 features (1 shape, 3 intensity, 5 texture). For clustering, the agglomerative hierarchical clustering algorithm grouped features into three clusters, selecting 21 features using a variance filter with a 0.02 threshold. Of these, 3 were intensity, 3 shape, and 15 texture features. Similarly, k-means clustering, with an optimal cluster count of 3, selected 21 features (4 intensity, 1 shape, and 15 texture features) using the same variance filter threshold.

For PET analysis, the research aimed to determine whether the original questions were addressed, specifically reducing 98 variables to a smaller set. After analysis, 26 features were selected for PCA and 17 for hierarchical clustering. All analyses identified texture-based features as the most significant. The objective was achieved by reducing the feature count to between 20 and 40, resulting in a 75% reduction in dimensionality, creating data matrices of 73 by 26 for PCA and 73 by 17 for hierarchical clustering.

For PCA, five main components captured 82.7% of the variance, with the remaining components contributing less than 11.5%. A loadings filter with a 0.15 threshold identified 26 features. Principal Component 1 contained 14 features (3 intensity, 1 shape, 10 texture), while Principal Component 2 had 1 feature, and Principal Component 3 had 2 shape features. Principal Component 4 had 5 texture features, and Principal Component 5 had 4 texture features. Texture-based features were concluded to be the most important according to PCA. For clustering analysis, 17 features were selected using a variance filter with a 0.05 threshold. Cluster 1 had 2 intensity, 1 shape, and 2 texture features; Cluster 2 had 1 intensity, 1 shape, and 4 texture features; Cluster 3 had 2 texture features; and Cluster 4 had 1 texture and 3 intensity features. The findings confirmed that texture features were the dominant contributors to the new data matrix.

FUTURE WORKS

An in-depth analysis using manifold learning techniques, such as ISOMAP and Laplacian eigenmaps, is critical for validating the adequacy of the selected features and confirming whether texture-based features are truly the most significant in cancer data analysis. These nonlinear dimensionality reduction methods provide a deeper understanding of the underlying data structure by preserving the geometry of high-dimensional spaces. This allows for a more robust assessment of feature importance, especially when conventional methods may overlook subtle patterns.

Given that cancer is one of the leading causes of death worldwide, with millions of lives affected each year, ensuring the accuracy and efficiency of diagnostic models is essential. Texture-based features have shown potential as key biomarkers in imaging studies, aiding in early detection, classification, and prognosis. By utilizing manifold techniques like ISOMAP and Laplacian, researchers can better evaluate whether these features provide the most meaningful contribution to the analysis, ensuring that no crucial data is lost during the dimensionality reduction process.

Furthermore, improving our ability to identify the most relevant features not only enhances diagnostic precision but also contributes to more personalized treatment strategies. This is crucial in the ongoing

fight against cancer, as early and accurate diagnosis significantly increases the chances of successful treatment. The more we refine our understanding of these complex diseases through advanced analytical techniques, the greater the strides we can make in reducing the human and economic burden cancer imposes on individuals, families, and healthcare systems worldwide.

CONTACT INFORMATION

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

Mostafa Zahed, East Tennessee State University, zahedm@etsu.edu

REFERENCES

1. An Introduction to Analysis, 2nd edition, by J. R. Kirkwood, Published by PWS Publishing Company and Waveland Press, Inc. 1995.
2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016;278(2):563-577.
3. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*. 2017;14(12):749-762.
4. Pesapane, F., Codari, M., & Sardanelli, F. (2018). Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, 2, Article 35. <https://doi.org/10.1186/s41747-018-0061-6>.
5. Parmar, C., Grossmann, P., Rietveld, D., Rietbergen, M. M., Lambin, P., & Aerts, H. J. (2015). Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Frontiers in Oncology*, 5, 272. <https://doi.org/10.3389/fonc.2015.00272>.
6. Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., ... & Gillies, R. J. (2018). Data for NSCLC radiogenomics collection. The Cancer Imaging Archive.
7. Dong, Y., Feng, Q., Yang, W., Lu, Z., Deng, C., Zhang, L., ... & Cui, C. (2016). Preoperative prediction of sentinel lymph node metastasis in breast cancer by radiomic signatures from dynamic contrast-enhanced MRI. *Journal of Magnetic Resonance Imaging*, 44(6), 1468-1477.
8. Hu, Y., Xie, C., Yang, H., Ho, J. W. K., Wen, J., Han, L. J., ... & Chen, W. (2019). Assessment of intratumoral heterogeneity and prognosis using CT textural features in advanced lung adenocarcinoma treated with epidermal growth factor receptor tyrosine kinase inhibitors. *American Journal of Roentgenology*, 213(3), W176-W184.
9. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9):1234-1248.
10. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
11. Yip, S. S. F., & Aerts, H. J. W. L. (2016). Applications and limitations of radiomics. *Physics in Medicine & Biology*, 61(13), R150-R166.
12. Vallières, M., Kay-Rivest, E., Perrin, L. J., Liem, X., Furstoss, C., Aerts, H. J. W. L., ... & Skamene, S. R. (2015). Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific Reports*, 7, 10117.
13. Rizzo S, Botta F, Raimondi S, et al. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp*. 2018;2(1):36.

14. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*. 2017;14(12):749-762.
15. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016;278(2):563-577.
16. P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou and H. Benali, "From Handcrafted to Deep-Learning-Based Cancer Radiomics: Challenges and Opportunities," in *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 132- 160, July 2019, doi: 10.1109/MSP.2019.2900993. keywords: Cancer;Feature extraction;Tumors;Predictive models;Medical diagnostic imaging;Deep learn- ing;Machine learning.
17. Zahed Mostafa and Skafyan Maryam, Application of feature selection and di- mension reduction techniques on large-scale ct dataset for lung cancer diagnosis based on radiomics-sesug 2022 paper 222, (2022), 5–13.
18. Jia Weikuan, Sun Meili, Lian Jian, and Hou Sujuan, Feature dimensionality reduction: a review, *Complex & Intelligent Systems* 8 (2022), 2198–6053.
19. Michael Berry and Azlinah Mohamed, Supervised and unsupervised learning for data science, 01 2020.
20. C Rencher Alvin, *Methods of multivariate analysis*, john wiley&sons inc, Publi- cation, Canada (2002).
21. Laurens van der Maaten and Geoffrey Hinton (2008), "Visualizing Data using t-SNE", in the *Journal of Machine Learning Research*.
22. Joshua B Tenenbaum, Vin de Silva, and John C Langford, A global geometric framework for nonlinear dimensionality reduction, *science* 290 (2000), no. 5500, 2319–2323.
23. Parnian Afshar, Arash Mohammadi, Konstantinos N Plataniotis, Anastasia Oikonomou, and Habib Benali, From handcrafted to deep-learning-based can- cer radiomics: challenges and opportunities, *IEEE Signal Processing Magazine* 36 (2019), no. 4, 132–160.
24. 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.
25. Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*.
26. 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).
27. Liang H., Sun X., Sun Y., and Gao Y., Text feature extraction based on deep learning: a review., *EURASIP Journal on Wireless Communications and Net- working* (2017), 211.
28. 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.
29. Faiella, E., Santucci, D., Calabrese, A., Russo, F., Vadal`a, G., Beomonte Zo- bel, B., Soda, P., Iannello, G., de Felice, C., & Denaro, V. (2022). Artificial intelligence in bone metastases: An MRI and CT imaging review. **Interna- tional Journal of Environmental Research and Public Health**, *19*(4), 2184. <https://doi.org/10.3390/ijerph19042184>
30. J MacQueen, Classification and analysis of multivariate observations, 5th Berke- ley Symp. Math. Statist. Probability (1967), 281–297.
31. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An intro- duction to statistical learning, vol. 112, Springer, 2013.

32. Edge, S.B., Compton, C.C. The American Joint Committee on Cancer:the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM . Ann Surg Oncol 17, 1471–1474 (2010).
33. Zhao, B. (2021). Understanding sources of variation to improve the re- producibility of radiomics. *Frontiers in Oncology*, *11*, Article 826214. <https://doi.org/10.3389/fonc.2021.826214>
34. Awadallah, M. (2018). Staging and treatment implications of the 7th vs. 8th AJCC staging system for oral cancer. Journal of Oral and Maxillofacial Surgery.
35. Visalakshi, S., & Radha, V. (2014). A literature review of feature selection tech- niques and applications: Review of feature selection in data mining.
36. Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). "Applied Logistic Regression" (3rd ed.). Wiley Series in Probability and Statistics.
37. Chen, B., Yang, L., Zhang, R., Luo, W., & Liu, W. (2020). Radiomics: An overview in lung cancer management—a narrative review. Annals of Transla- tional Medicine, 8(11), 712.