Main Objective of the Analysis

The main objective of this analysis is to apply unsupervised learning techniques to a dataset of lung cancer imaging to identify intrinsic groupings that may correlate with various stages or types of lung cancer. Through dimensionality reduction, we aim to uncover patterns that are not immediately apparent in the high-dimensional raw data. The benefits of this analysis to stakeholders include the potential to streamline diagnostic processes and tailor treatment strategies to specific cancer subtypes.

Description of the Data Set

The dataset comprises 475 lung cancer images, each originally varying in size and some in color. Sourced from a public medical image repository, these images represent a wide range of cases, from early to advanced stages of cancer. Our goal is to apply clustering to these images to see if natural groupings emerge that could correlate with known cancer stages or other clinical parameters.

Data Exploration and Preprocessing

Initial exploration revealed a mix of grayscale and RGB images of varying sizes. To standardize the dataset, each image was converted to grayscale and resized to 512x512 pixels. Pixel values were normalized to a 0-1 range. Images were then flattened into vectors to prepare for analysis, resulting in each image being represented by a 262,144-dimensional vector.

Unsupervised Learning Model Variations

We trained three unsupervised models:

PCA for dimensionality reduction, where we reduced the feature space while retaining 95% of the variance.

DBSCAN to identify clusters based on density, with an epsilon value of 0.5 and a minimum sample count of 5.

Hierarchical Clustering to discern the hierarchical grouping of data points, utilizing Ward's method.

Each model was selected based on its ability to handle high-dimensional data and to provide different perspectives on the inherent groupings within the dataset.

Model Selection

DBSCAN was chosen as the most suitable model due to its capability to work with the complexity of medical imaging data. It effectively highlighted distinct groups and isolated anomalies, which could correlate with unusual or aggressive cancer manifestations.

Key Findings and Insights

DBSCAN's application unveiled three significant clusters and various outliers within the PCA-reduced space. Some clusters seemed to correlate with certain stages of cancer, suggesting a possible link between image features and pathological stages. Outliers identified by DBSCAN may correspond to rare or atypical cases, warranting further medical investigation.

Next Steps and Improvements

Further steps involve the integration of patient metadata to contextualize clusters and validate findings clinically. Future work could also explore supervised learning for predictive modeling, with

the current clusters serving as labels. To improve the model, it's recommended to test different epsilon values and minimum sample sizes for DBSCAN to refine the clustering granularity.