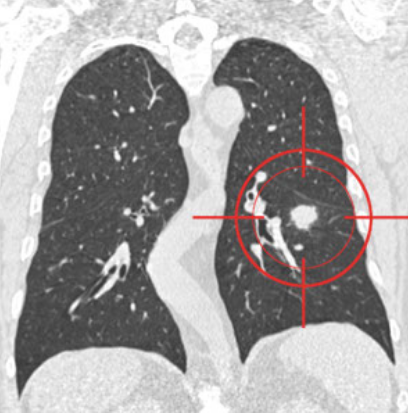


**ML Challenge 2 Milestone 2 Report**



Aleix Ramon Brugada Iglesias, 41655484T

Adrián Redondo Fernández, 48181224L

María Fernández Barrera, 46837717B

Soufiane Aatab, X6132088N

Júlia Casanovas i Saumell, 26071980P

* 24-25 MACHINE LEARNING (150400)

16/05/2025

**1. Introduction**

In Milestone 2, we delve deeper into the radiological annotations in file ‘MetadatabyNoduleMaxVoting.xlsx’ to uncover underlying structure in lung nodule annotations. Building on our initial data preparation, we leverage unsupervised learning to group nodules by texture, shape, and margin characteristics, aiming to highlight patterns that may correlate with malignancy.

**1.1 Goals**

Perform exploratory clustering to identify natural groupings in the radiological annotations feature data. Use statistical tests to determine which features drive cluster separation. Compare clustering stability across K-Means and hierarchical methods.

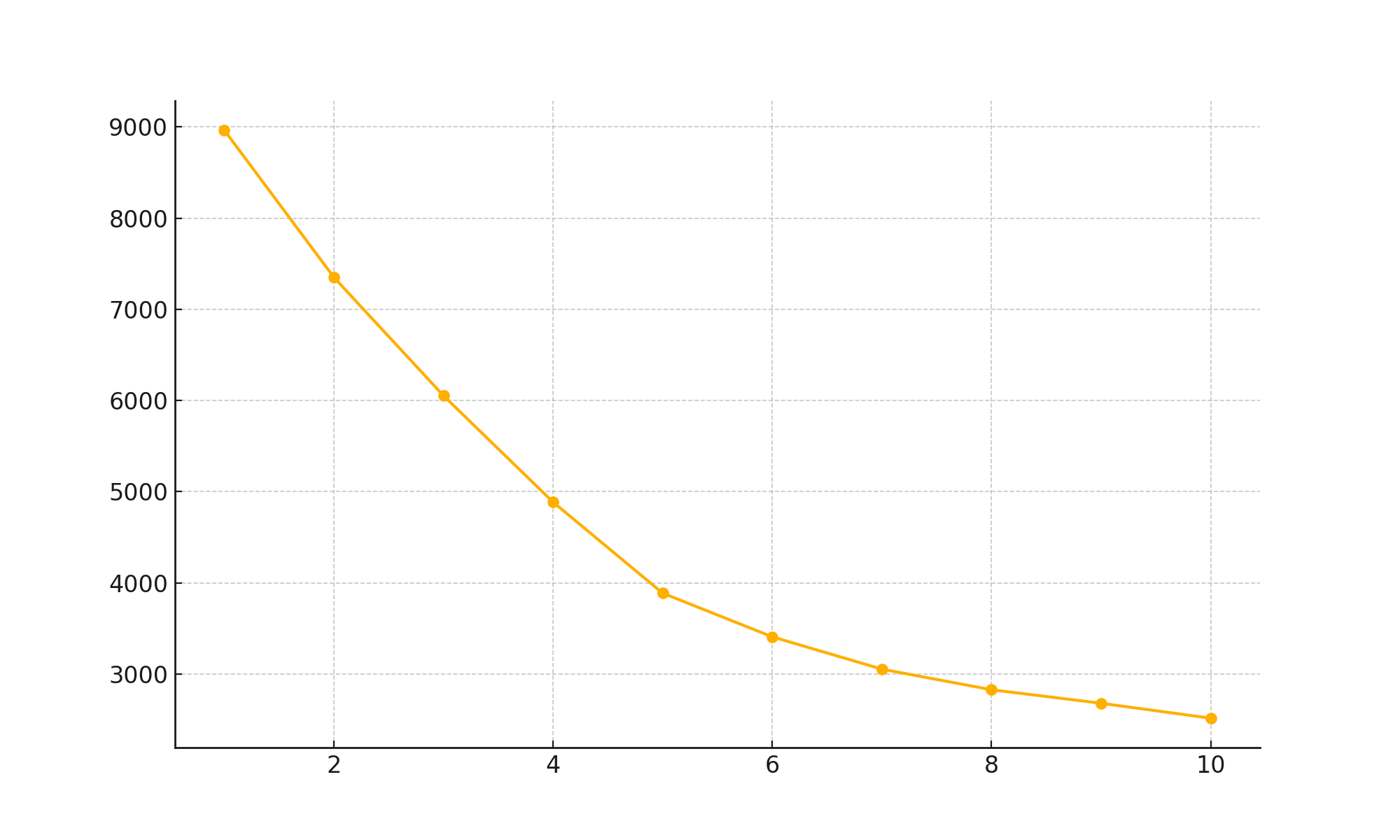
**2. Methodology**

Our methodology involves four key steps: preprocessing the annotation features, applying two clustering techniques, reducing dimensionality for visualization, and conducting statistical evaluations to interpret cluster differences.

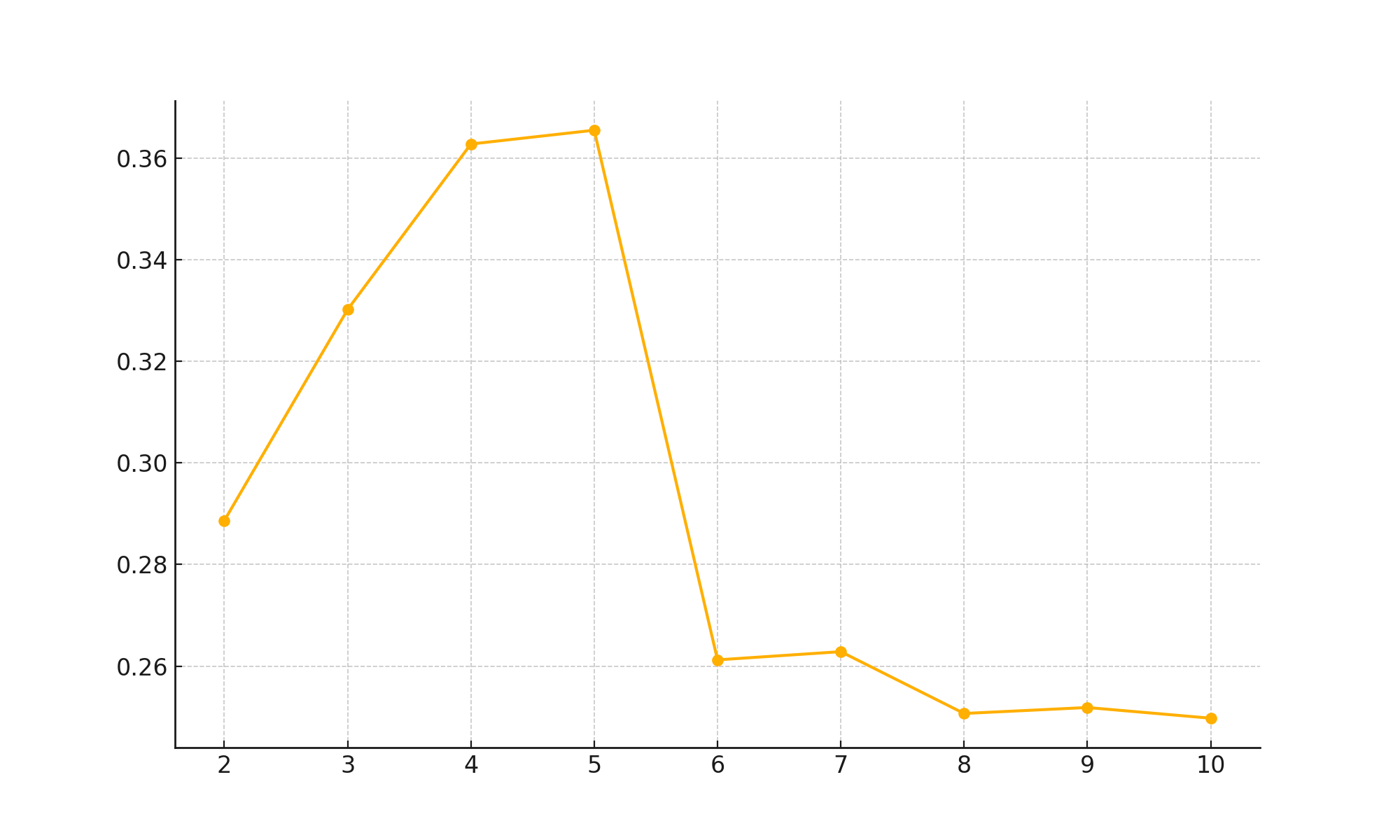
**2.1 K-Means Clustering**

We applied K-Means over K=1 to 10 clusters, evaluating inertia and silhouette metrics to select the optimal cluster count.

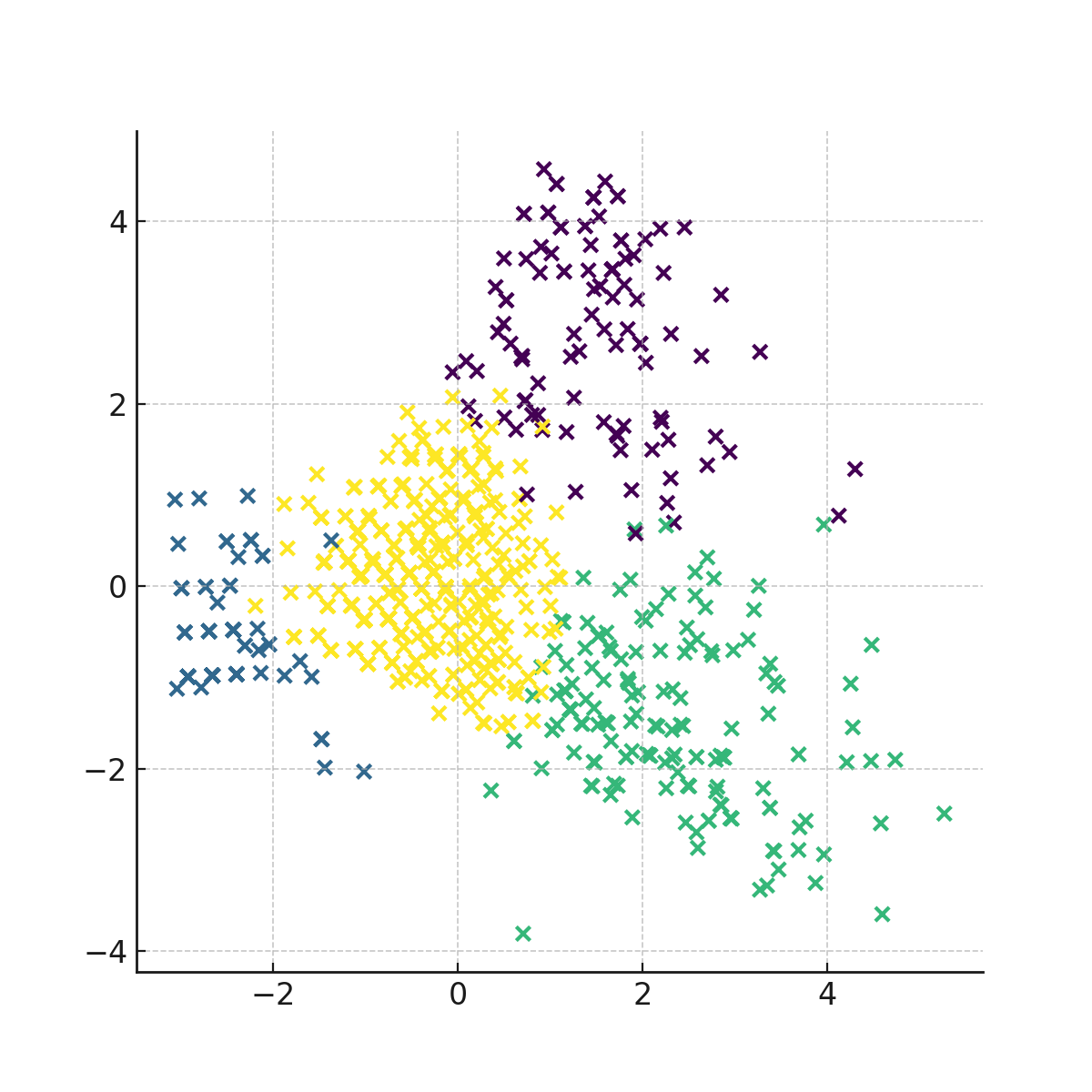
*(Figure 1: Elbow plot showing inertia vs. number of clusters)*



*Figure 2: Silhouette scores for K from 2 to 10)*



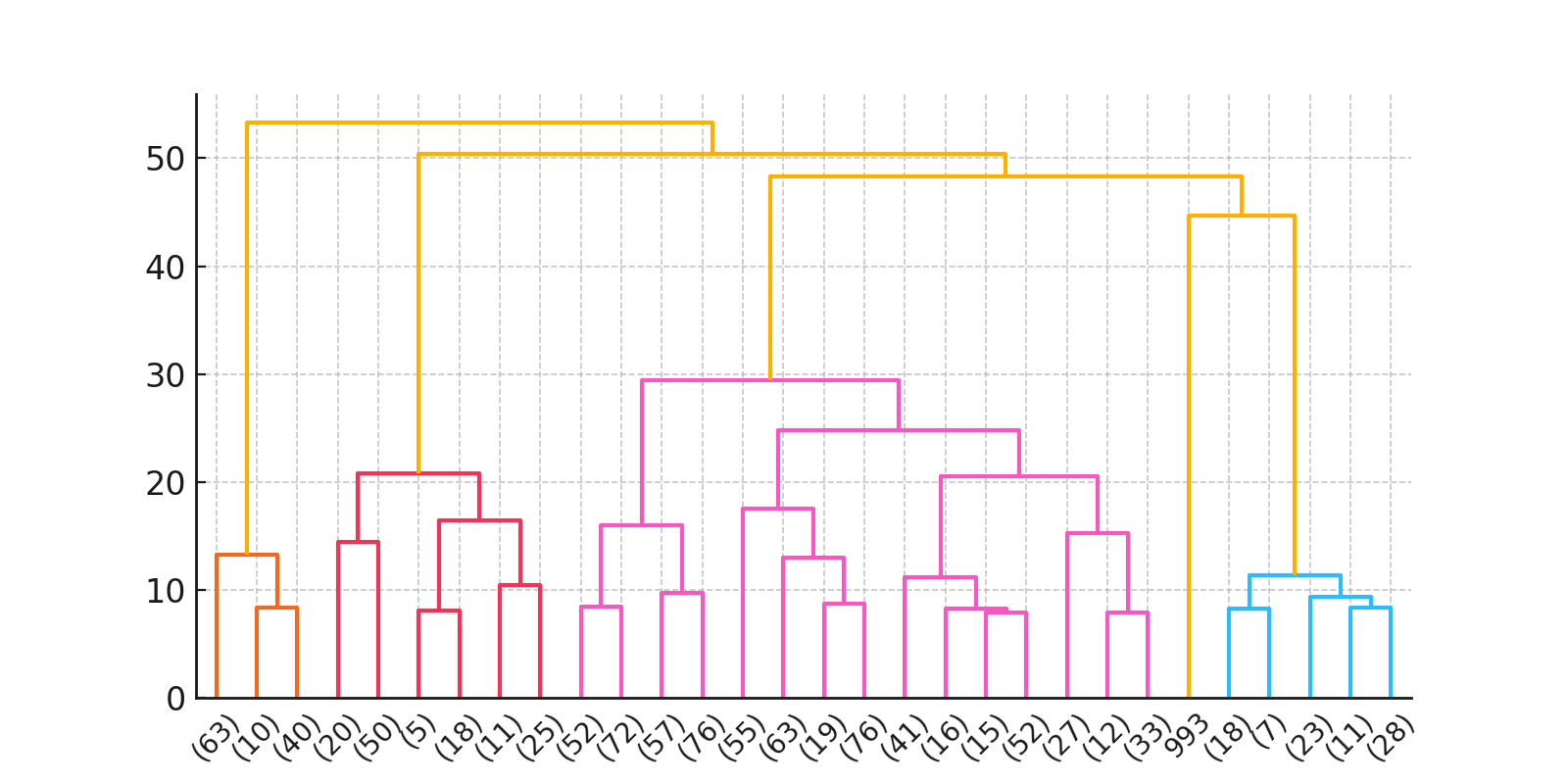
*(Figure 3: PCA projection of data colored by K-Means cluster assignments)*



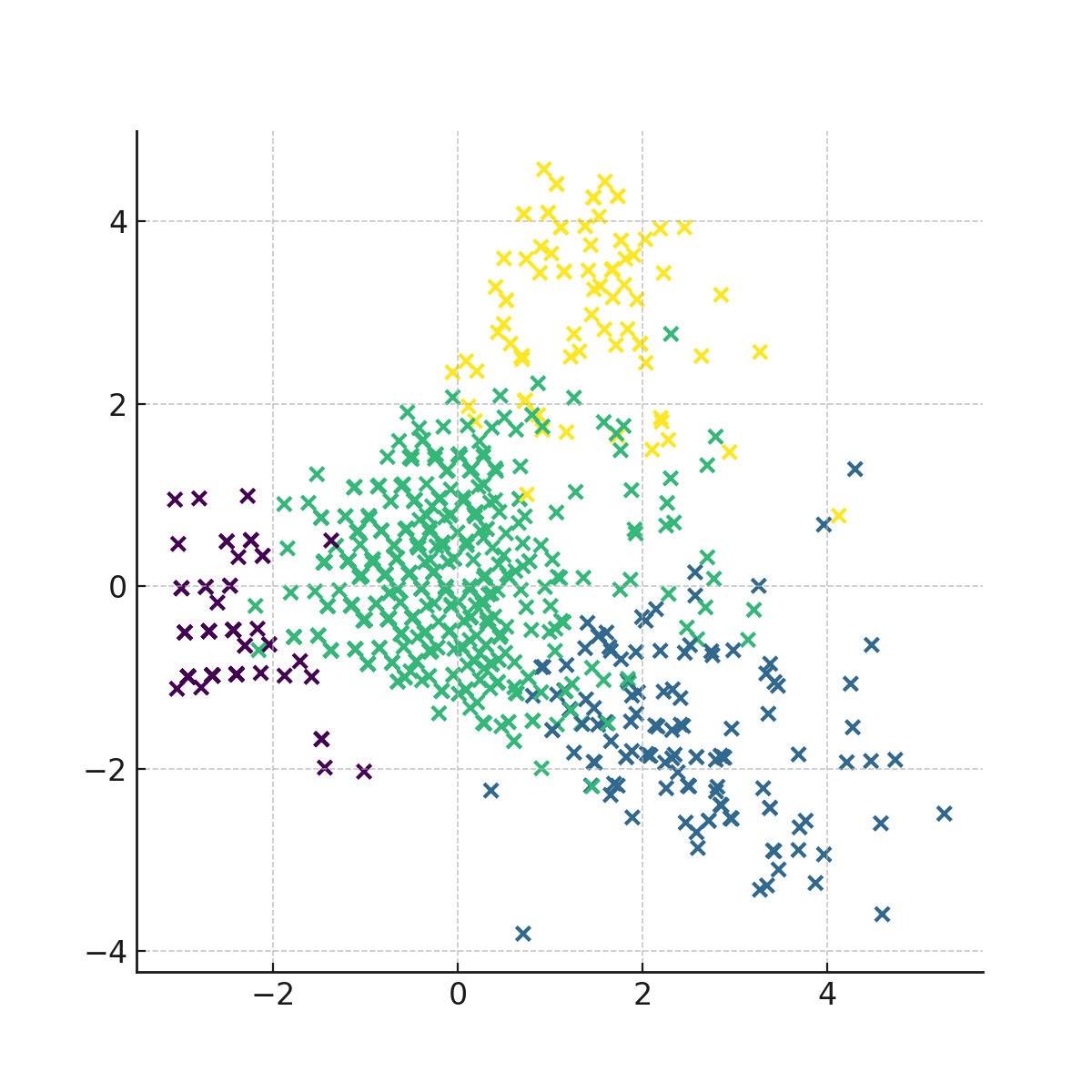
**2.2 Hierarchical Clustering**

We used Ward’s linkage to build a dendrogram and cut at 4 clusters, paralleling our K-Means choice for comparison.

*(Figure 4: Hierarchical clustering dendrogram with Ward’s linkage)*



*(Figure 5: PCA projection of data colored by hierarchical cluster assignments)*



**3. Experimental Design**

**3.1 Dataset Description**

We utilized the LUNA16 dataset metadata, where each nodule’s features are aggregated by max-voting across expert annotations. This yielded 996 samples described by nine quantitative features.

**3.2 Statistical Metrics**

Cluster outcomes were assessed by one-way ANOVA to highlight features with significant between-group variance, and correlation heatmaps were generated to examine inter-feature relationships conditional on cluster membership.

**4. Results**

**4.1 ANOVA on Clusters**

One-way ANOVA revealed that features such as calcification, margin, and spiculation show the strongest differences across K-Means clusters, with p-values < 0.001 (Table 1). These results indicate which descriptors drive cluster formation.

*(Table 1: K-means ANOVA results)*

| Feature | F-statistic | p-value |
| --- | --- | --- |
| Malignancy\_value | 361.22 | 0.0000 |
| Calcification\_value | 8672.90 | 0.0000 |
| InternalStructure\_value | 2.78 | 0.0398 |
| Lobulation\_value | 238.77 | 0.0000 |
| Margin\_value | 284.36 | 0.0000 |
| Sphericity\_value | 25.48 | 0.0000 |
| Spiculation\_value | 356.30 | 0.0000 |
| Subtlety\_value | 114.57 | 0.0000 |
| Texture\_value | 2275.04 | 0.0000 |

One-way ANOVA on hierarchical clusters shows that calcification, texture, and spiculation exhibit the largest between-group differences (p < 0.001; Table 2), indicating these high-contrast features predominantly drive the hierarchy.

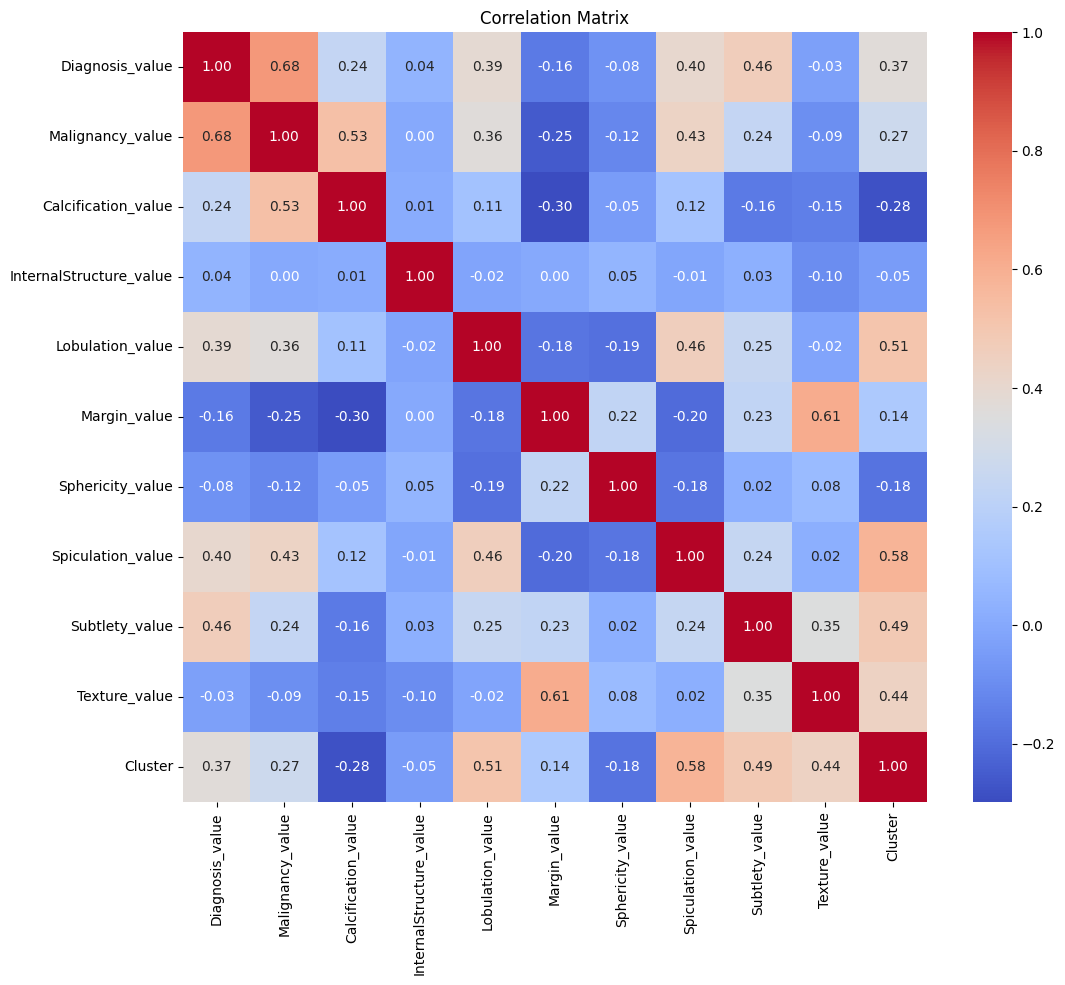
*(Table 2: HC ANOVA results)*

| Feature | F-statistic | p-value |
| --- | --- | --- |
| Malignancy\_value | 296.61 | 0.0000 |
| Calcification\_value | 9524.23 | 0.0000 |
| InternalStructure\_value | 3.48 | 0.0158 |
| Lobulation\_value | 198.10 | 0.0000 |
| Margin\_value | 207.01 | 0.0000 |
| Sphericity\_value | 13.56 | 0.0000 |
| Spiculation\_value | 445.36 | 0.0000 |
| Subtlety\_value | 111.35 | 0.0000 |
| Texture\_value | 1292.61 | 0.0000 |

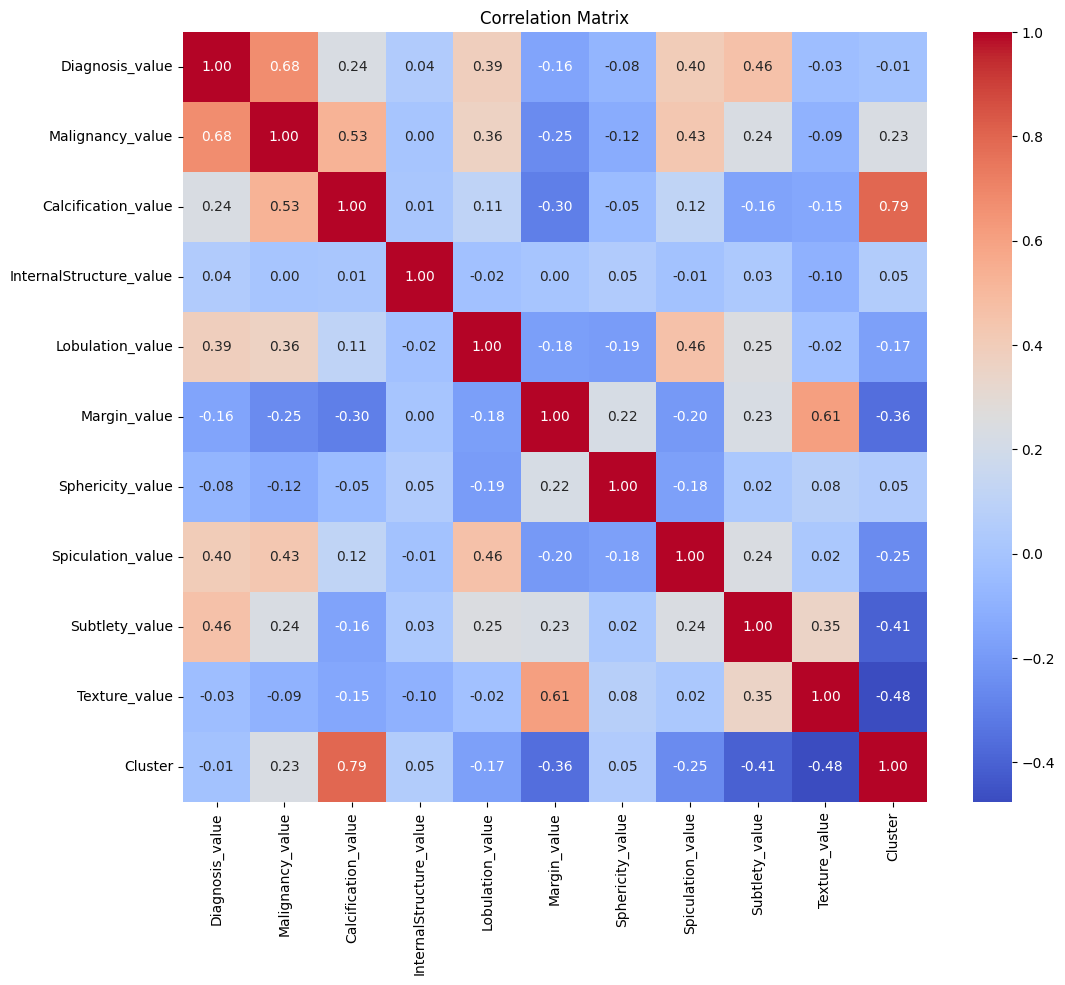
**4.2 Correlation Analysis**

Correlation heatmaps (Figures 6 and 7) show that texture and subtlety features correlate strongly within clusters, while the addition of cluster labels highlights how groups separate in feature–feature space.

*(Figure 6: Correlation matrix for K-Means clusters)*



*(Figure 7: Correlation matrix for hierarchical clusters)*



**5. Discussion and Conclusions**

The one-way ANOVA analyses (see Table 1 for K-Means and Table 2 for Hierarchical clustering) consistently identify **calcification**, **texture**, and **spiculation** as the most discriminative features driving cluster separation, while **internal structure** contributes minimally in both methods. K-Means produces a more balanced distribution of feature importance—malignancy value and morphological cues (margin, lobulation) also emerge with notable between-group differences—whereas Hierarchical clustering accentuates calcification as the dominant separator, downplaying some textural descriptors relative to K-Means.

The correlation heatmaps (Figures 6 and 7) corroborate these findings: K-Means clusters reflect a blend of texture and shape characteristics, while Hierarchical clusters align more strongly with high-contrast features like calcification. This suggests that K-Means may capture multifaceted radiomic patterns useful for nuanced subgrouping, whereas Hierarchical clustering isolates the single feature with the greatest variance.

In summary, features related to calcification and texture (particularly spiculation) are key drivers of the unsupervised grouping of nodules. Morphological features (margin, lobulation) play a supportive role, and less discriminative descriptors can be deprioritized. Moving forward, integrating these high-impact features into a supervised malignancy classifier is likely to yield the greatest predictive benefit.

**6. References**

[1] Python Software Foundation. Python Language Reference, version 3.x.

[2] Pedregosa et al. Scikit-learn: Machine Learning in Python, JMLR, 2011.

[3] van Griethuysen JJ et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res, 2017.

[4] <https://luna16.grand-challenge.org/Description/> (accessed May 2025).