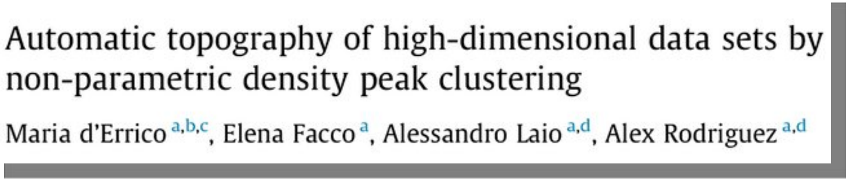
Unsupervised Learning Final Project

Implementation of Density Peaks Advanced Clustering on Breast Cancer Dataset

Author: Gabriel Gustavo Costanzo

Date: 28/06/2024



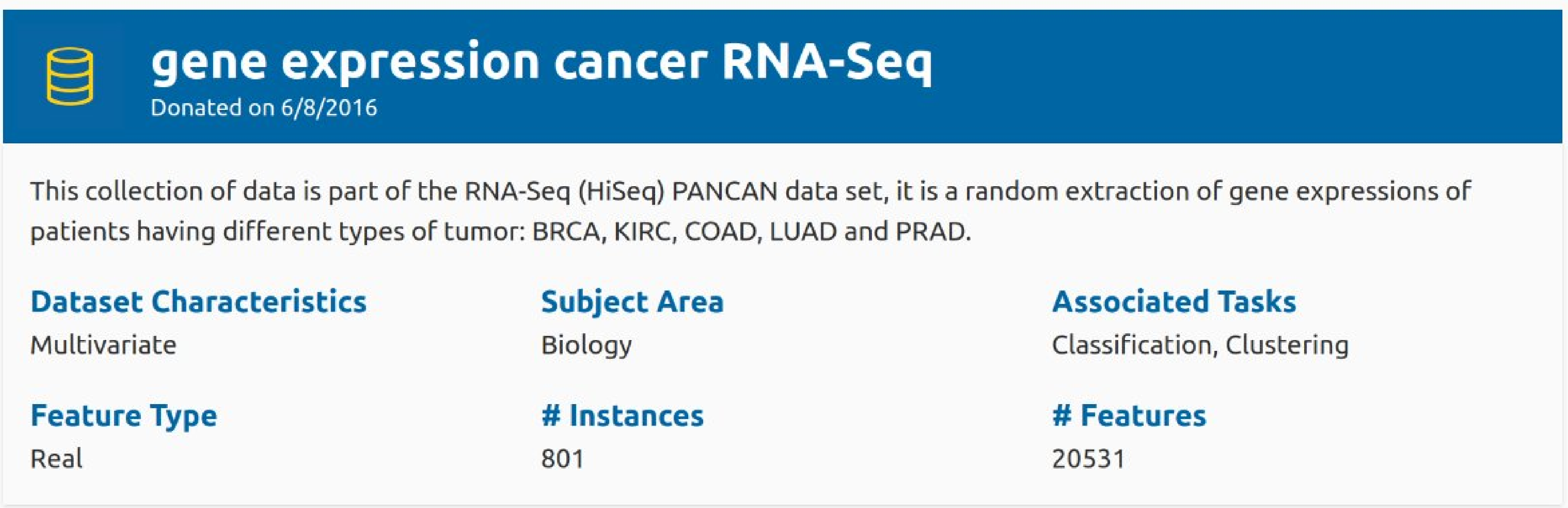
Automatic topography of high-dimensional data sets by non-parametric density peak clustering

Maria d'Errico [a,b,c], Elena Facco [a], Alessandro Laio [a,d], Alex Rodriguez [a,d]

Objective

Apply the Density Peaks Advanced (DPA) clustering algorithm to the Breast Cancer dataset to test its performance and ability to discriminate different types of cancer cells, compared with other clustering algorithms, with and without feature selection.

Data set



gene expression cancer RNA-Seq

Donated on 6/8/2016

This collection of data is part of the RNA-Seq (HiSeq) PANCAN data set, it is a random extraction of gene expressions of patients having different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Biology	Classification, Clustering
Feature Type	**# Instances**	**# Features**
Real	801	20531

COAD: Colon Adenocarcinoma
BRCA: Breast Invasive Carcinoma
LUAD: Lung Adenocarcinoma
PRAD: Prostate Adenocarcinoma
KIRC: Kidney Renal Clear Cell Carcinoma

Methology

1. Install DPA library
https://github.com/mariaderrico/DPA

2. Data preprocessing

3. Intrinsic dimension calculation.

4. Dimensionality reduction and visualization of high-dimensional gene expression.

5. Baseline clustering algorithm.

6. Selection of the best values for clustering algorithms parameters.

7. Subset feature selection, repeat point 5.

8. Results: Graphical representation of the clustering.

9. Results: Metrics evaluation.

10. Results: Topography analysis.

11. Conclusion

Data preprocessing

✓ Log(1 + x) transformation

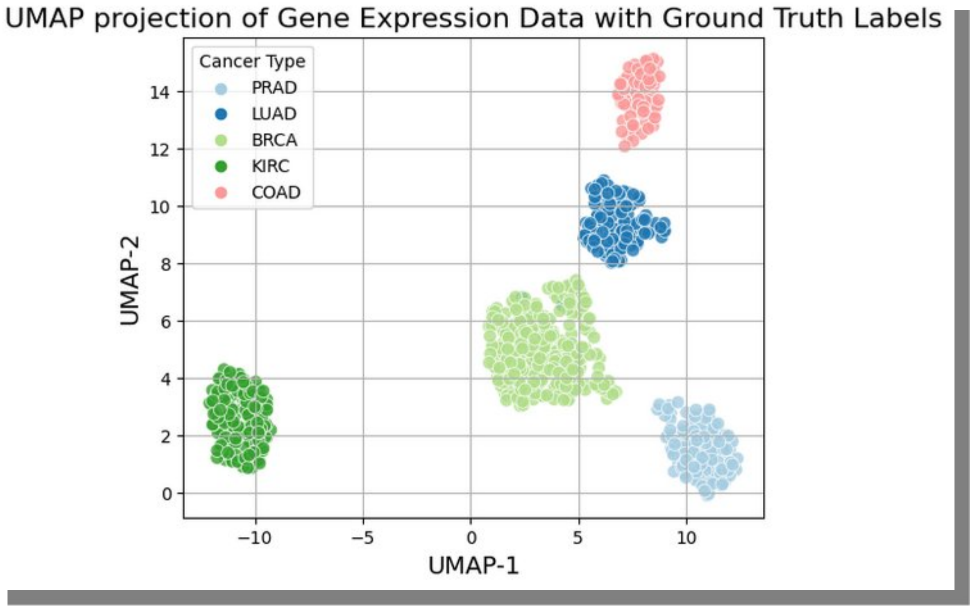
✓ Z-score normalization

✓ Handling of non-informative columns

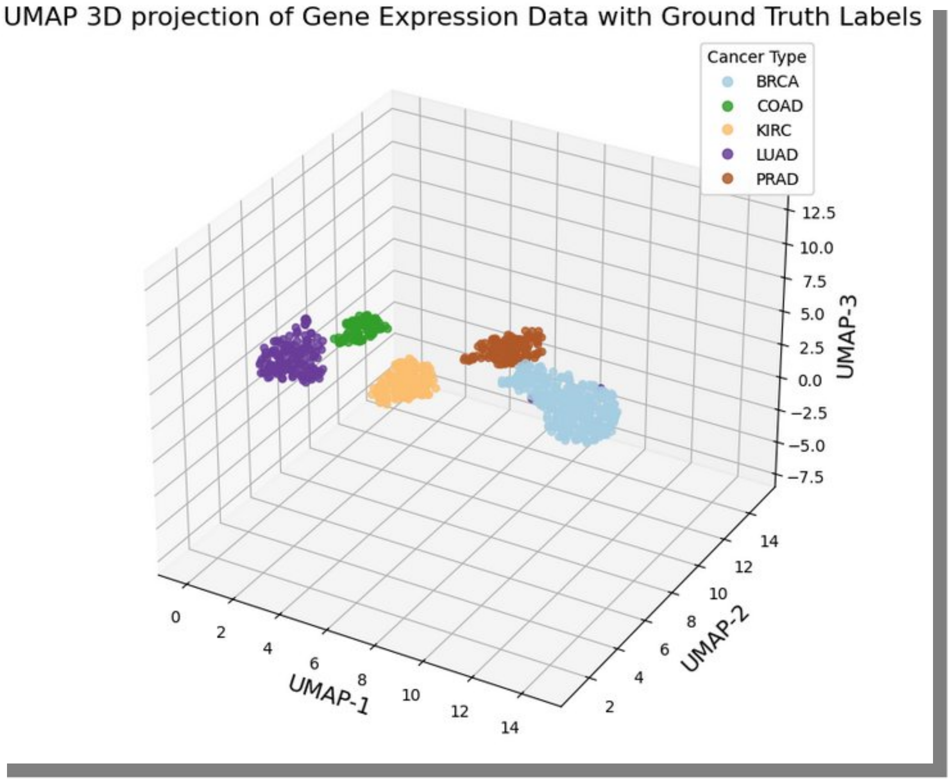
- After preprocessing remain 20,264 genes or features
- Intrinsic dimension calculation with TwoNN:
  38

Dimensionality reduction and visualization

using UMAP(Uniform Manifold Approximation and Projection)



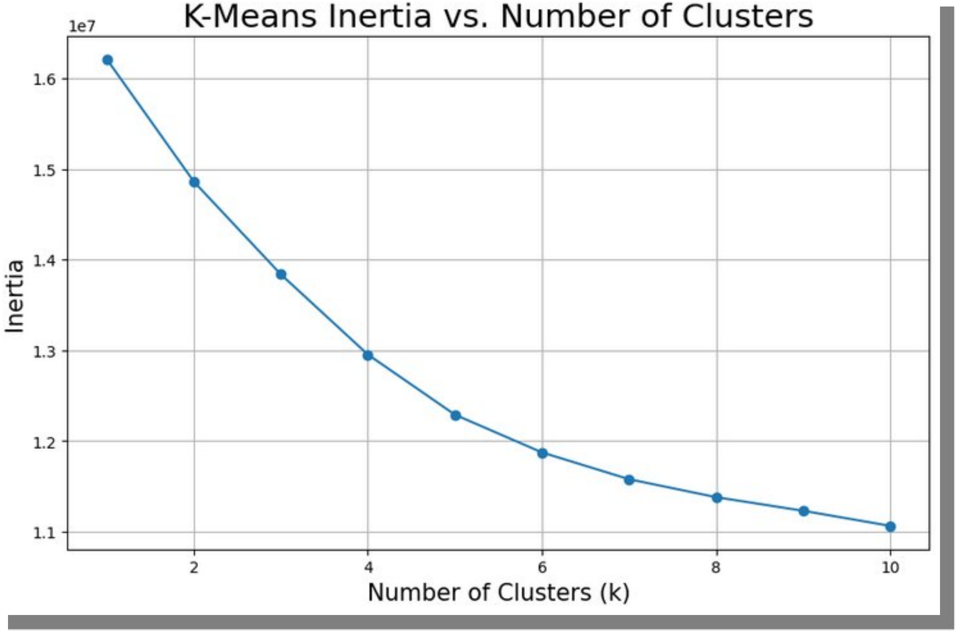
UMAP projection of Gene Expression Data with Ground Truth Labels

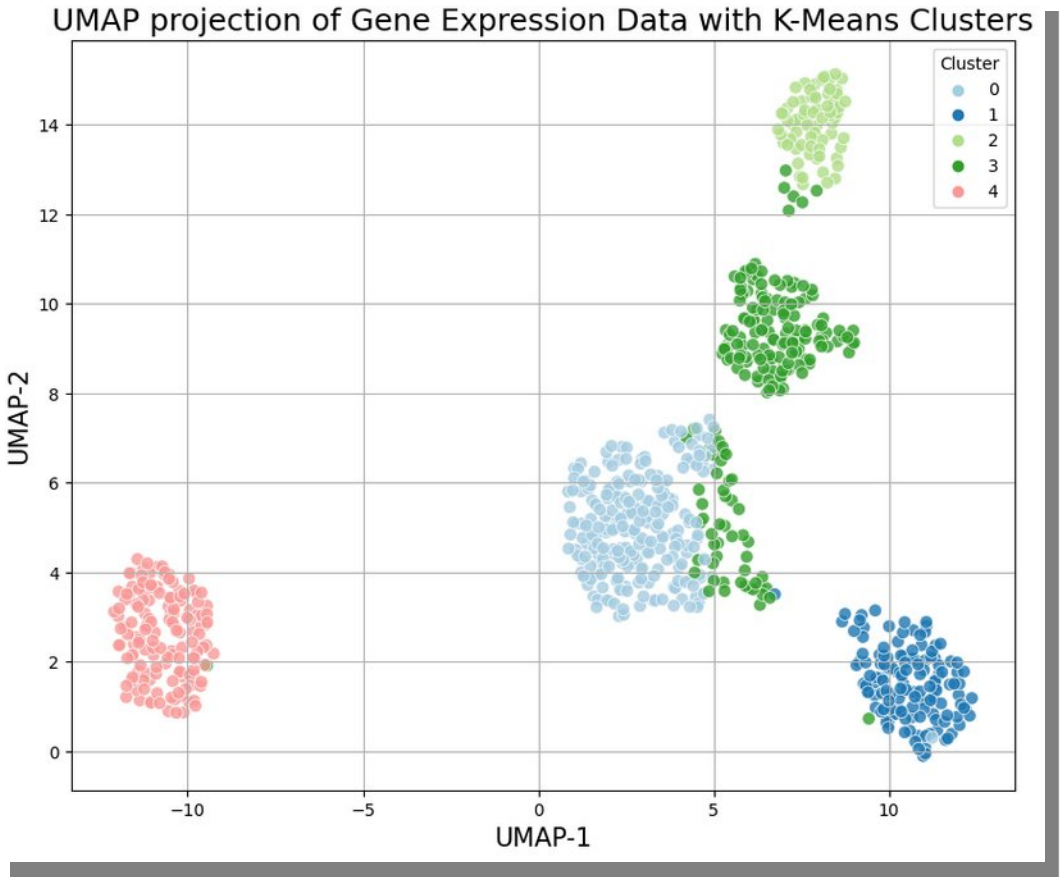


UMAP 3D projection of Gene Expression Data with Ground Truth Labels

K-means

Given the globular shape of the data groups or natural clusters and domain knowledge I chose K-means as a baseline clustering method.





Clustering algorithms

- ✓ Flat clustering: K-means

Non-Parametric Density Based Clustering: DPA and HDBSCAN

- ✓ Spectral clustering

Metrics

- Ajusted Rand Index. Range from -0.5 to 1

ARI = (RI - E) / (max(RI) – E), $R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$

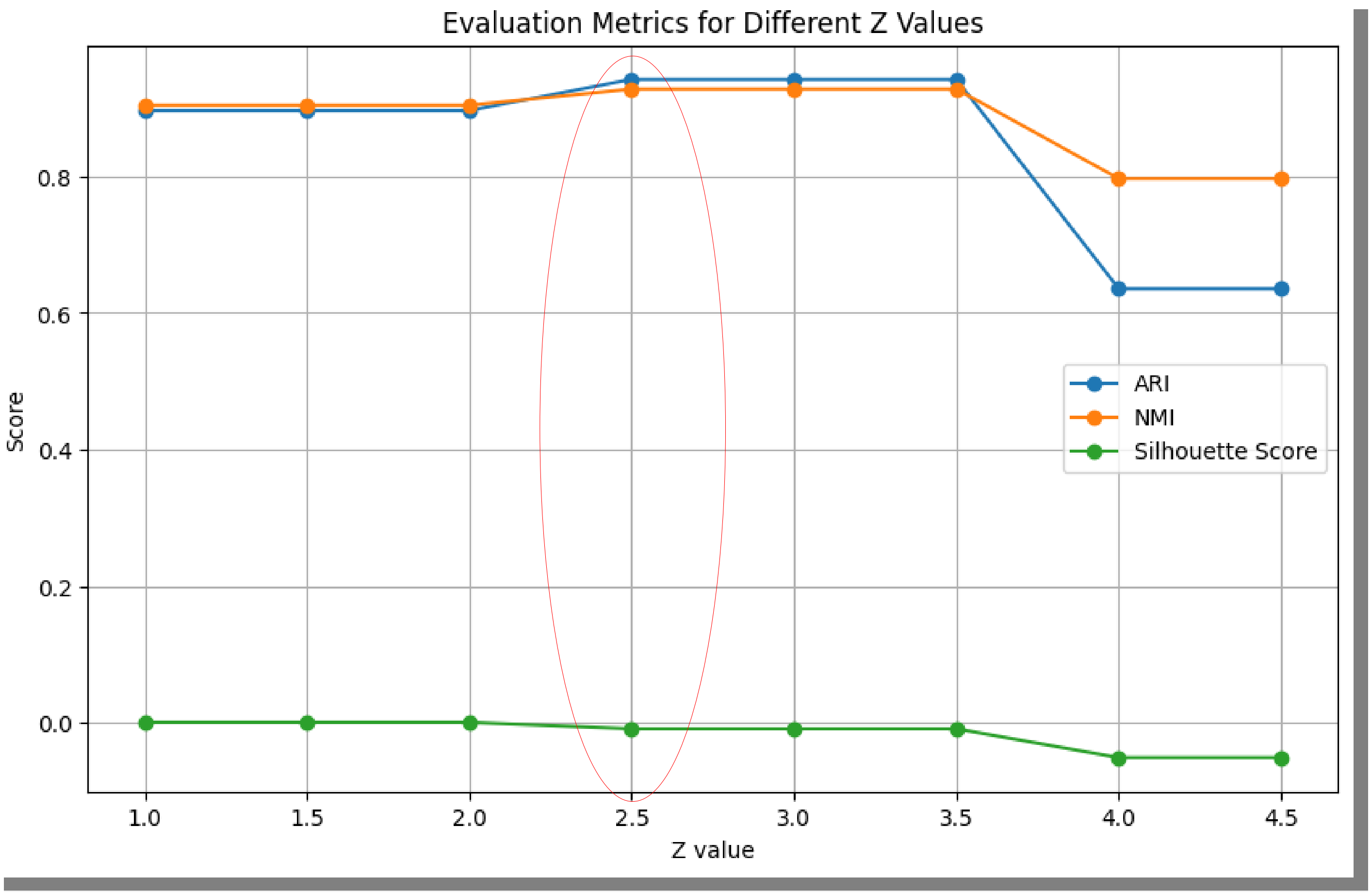
- Normalized Mutual Information (NMI). Range from 0 to 1

- Silhouette score. Range from -1 to 1.

  silhouette coefficient = (separation — cohesion) / max(separation, cohesion)

Selection of the parameter’s values for the clustering algorithms

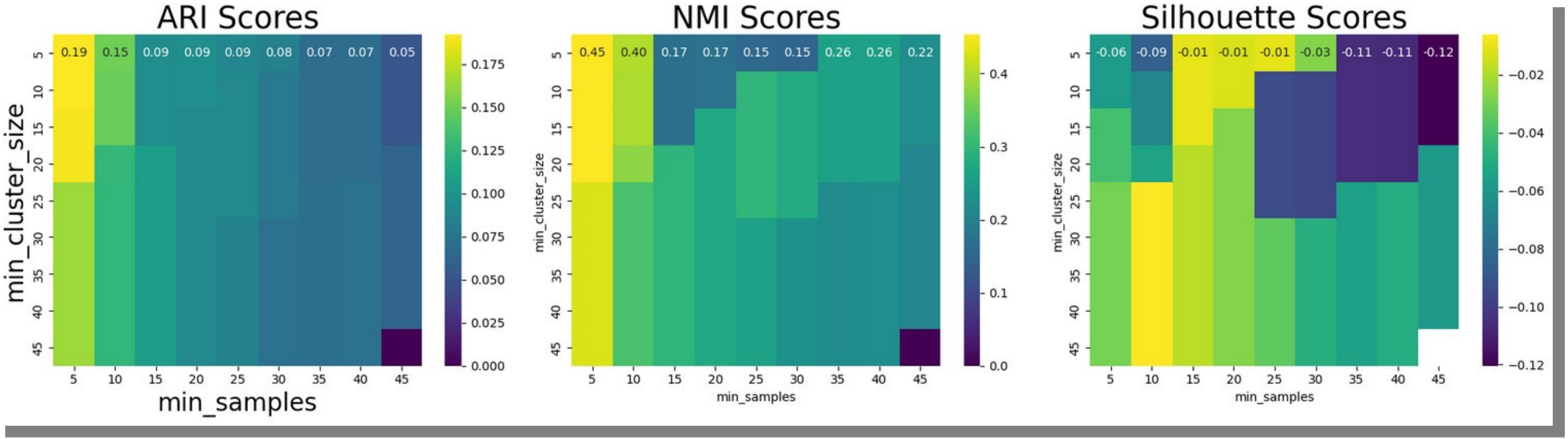
Density Peaks Advanced



Z determines the confidence level for distinguishing genuine density peaks from statistical fluctuations.

The best value for Z given the metrics should be 2.5 but, since the intrinsic dimension of the data set is 38, the error in the estimated density will be high, so I must choose a lower confidence Z=1.5

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)



min_samples= 5

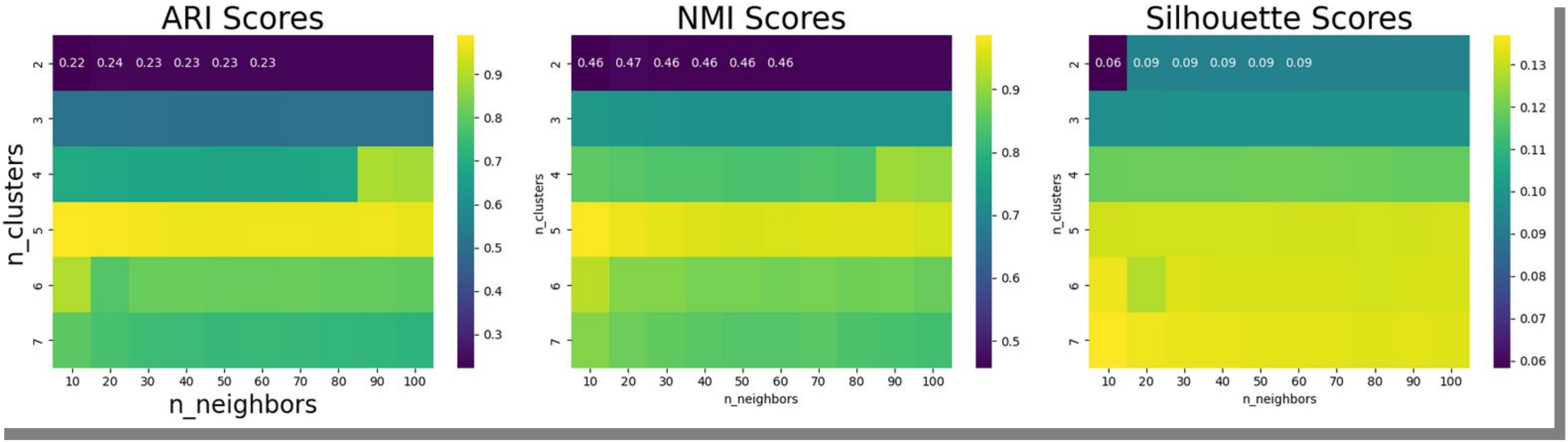
min_cluster_size= 20

ARI = 0.18914821

NMI = 0.45081652

Silhouette Score = -0.04027180

Spectral Clustering



n_neighbors= 10

n_clusters= 5

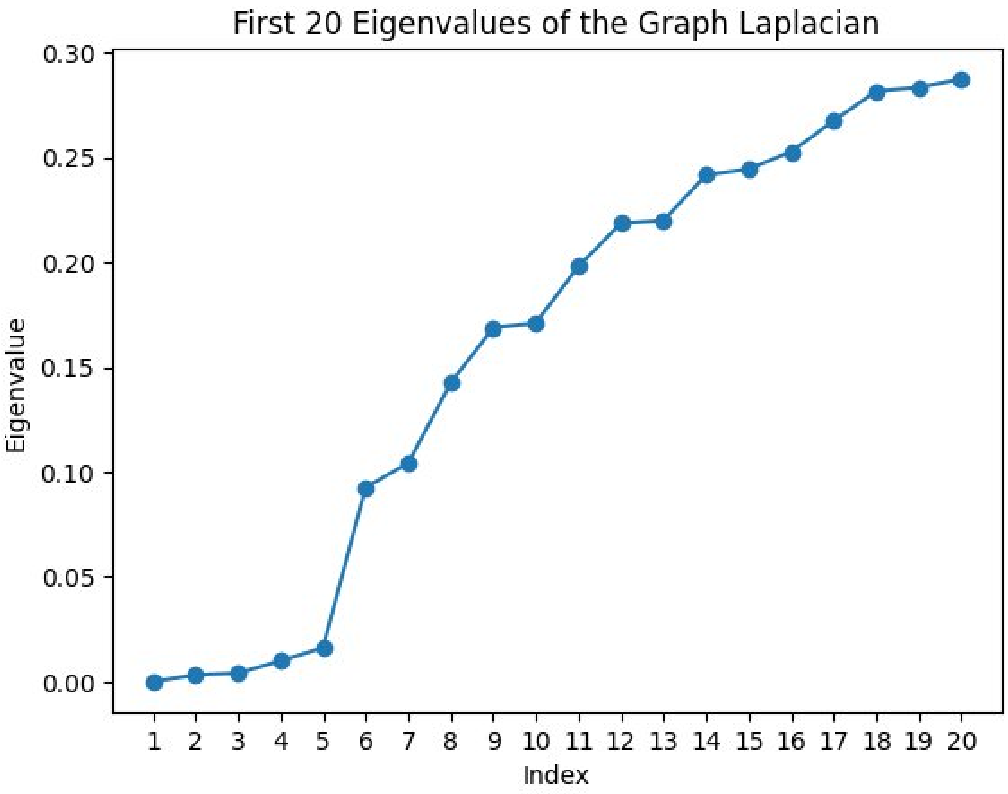
ARI: 0.98932736

NMI: 0.98570145

Silhouette Score: 0.13113107

Spectral Clustering

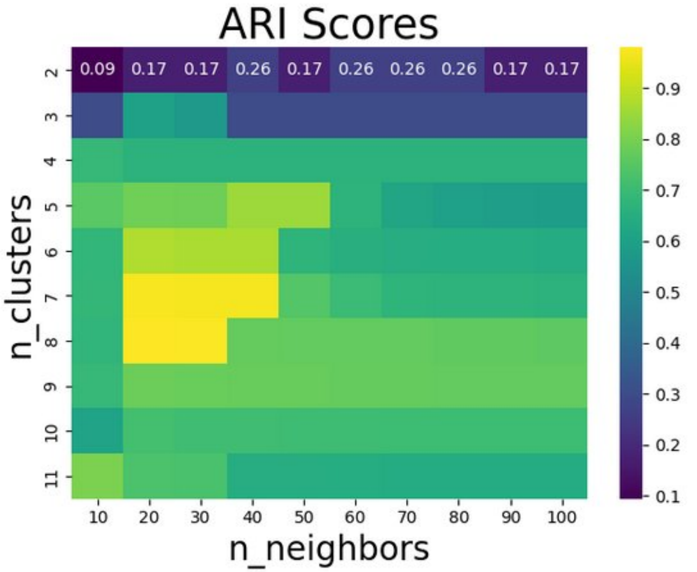
Determine the number of clusters using Eigenvalue Gaps.

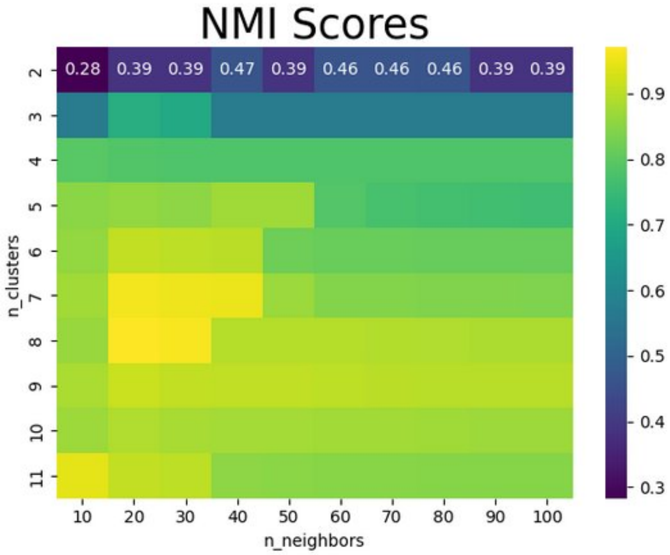


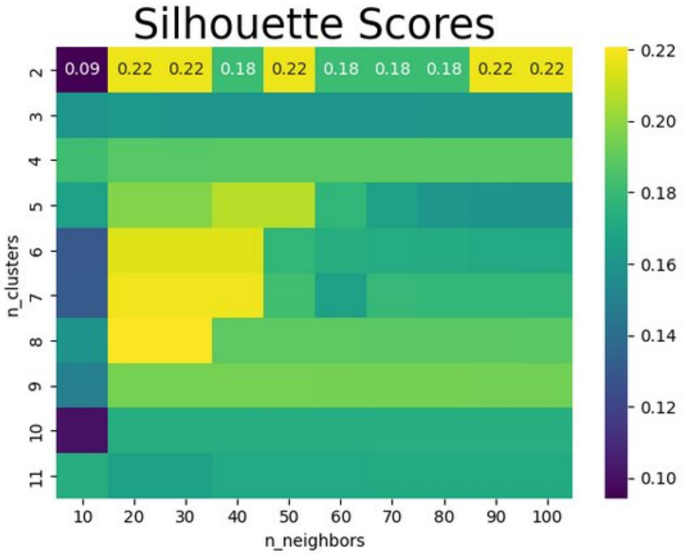
The max gap is between the 5th and 6th eigenvalue so, it suggest that the number of clusters is 5

Spectral Clustering

MNIST digit example (10 digits form 0 to 9)


ARI Scores


NMI Scores


Silhouette Scores

n_neighbors= 20

n_clusters= 8

ARI: 0.98932736

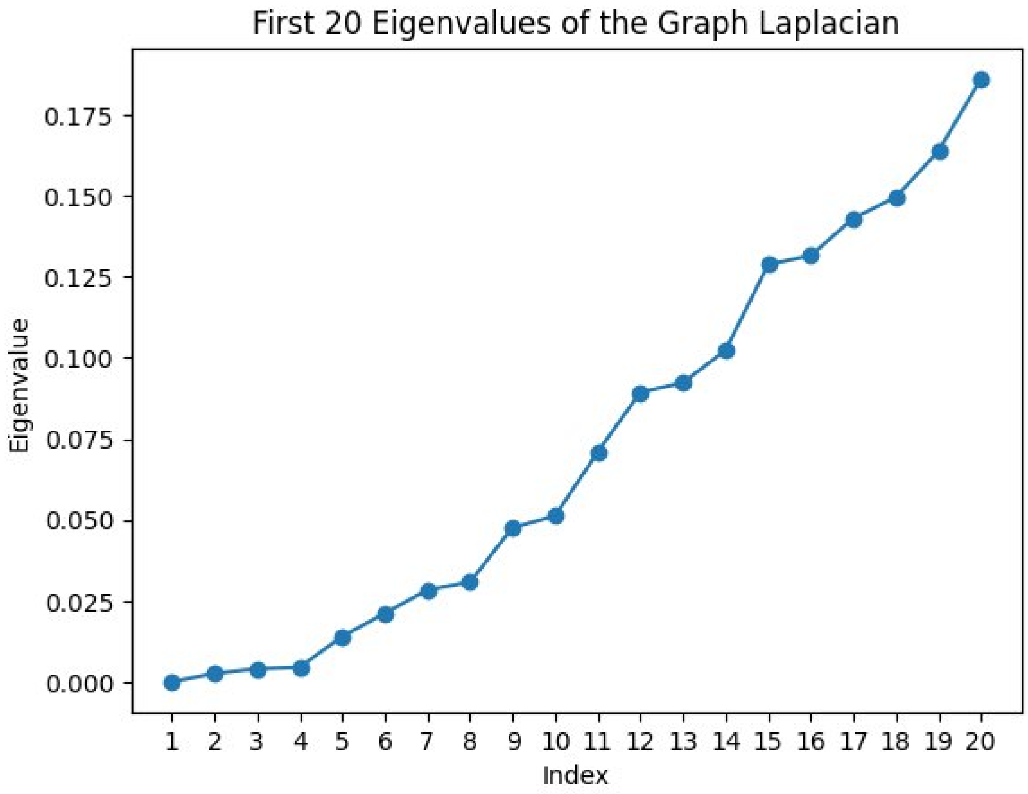
NMI: 0.98570145

Silhouette Score: 0.13113107

Spectral Clustering

MNIST digit example

Determine the number of clusters using Eigenvalue Gaps.



The max gap is between the 14th and 15th eigenvalue so, it suggest that the number of clusters is 14

Subset selection

Objective: To identify the top 1000 most relevant genes for cancer classification using feature importance from three supervised learning models and evaluate their performance.

1- Train Models:
Random Forest (RF): Trained to calculate feature importance.

Support Vector Machine (SVM): Linear kernel used to derive feature importance.

Logistic Regression (LR): Coefficients used to determine feature importance.

2- Compute Feature Importances:
Feature importances are obtained from RF, SVM, and LR models.

Calculate the average importance score for each gene.

3- Rank and Select Top Genes: Rank genes based on their average importance scores.
Select the top 1000 most important genes.

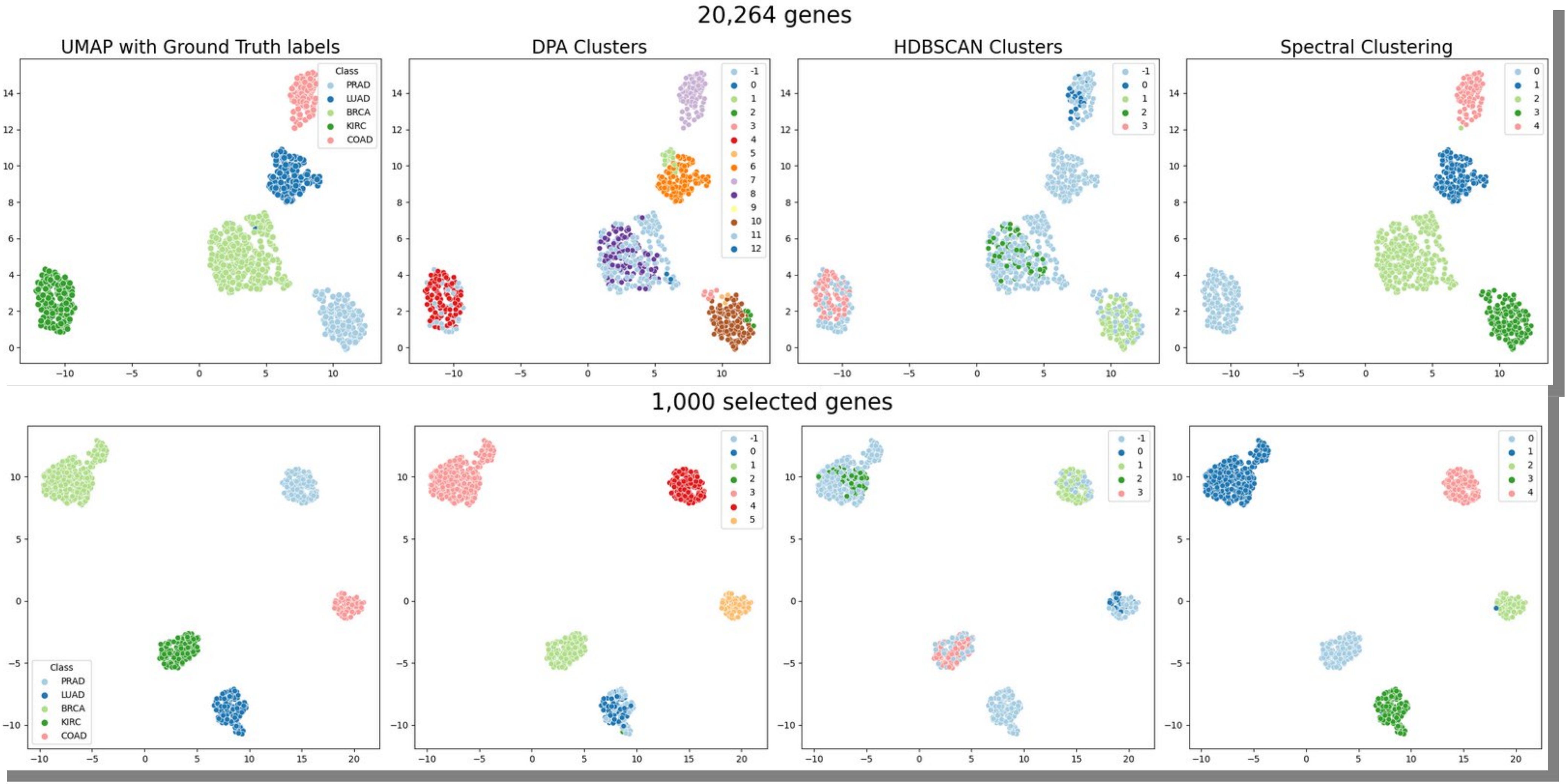
4- Evaluate Performance:

Use cross-validation to evaluate the classification accuracy of RF, SVM, and LR models on the subset of top 1000 genes.

ID= 32 and Z= 1.5

ID= 32 and Z= 1.5

Results: Clustering

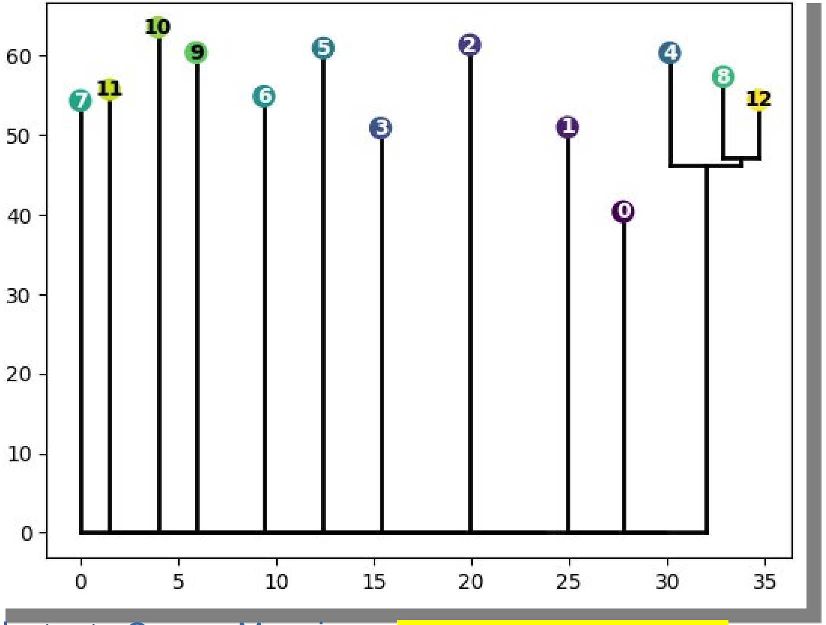


Results: Evaluation of metrics

		Original features				Subset feature			
Clustering algorithm./ Metrics	K-means	DPA	DPA_Halo	HDBSCAN	Spectral Clustering	DPA	DPA_Halo	HDBSCAN	Spectral Clustering
ARI	**0.79**	**0.94**	**0.64**	**0.19**	**0.99**	**1.0**	**0.95**	**0.92**	**1.0**
NMI	**0.85**	**0.93**	**0.79**	**0.45**	**0.99**	**0.93**	**0.79**	**0.91**	**1.0**
Silhouette Score	**0.13**	**-0.01**	**-0.02**	**-0.04**	**0.13**	**-0.01**	**-0.02**	**0.27**	**0.28**

Results: Topography analysis

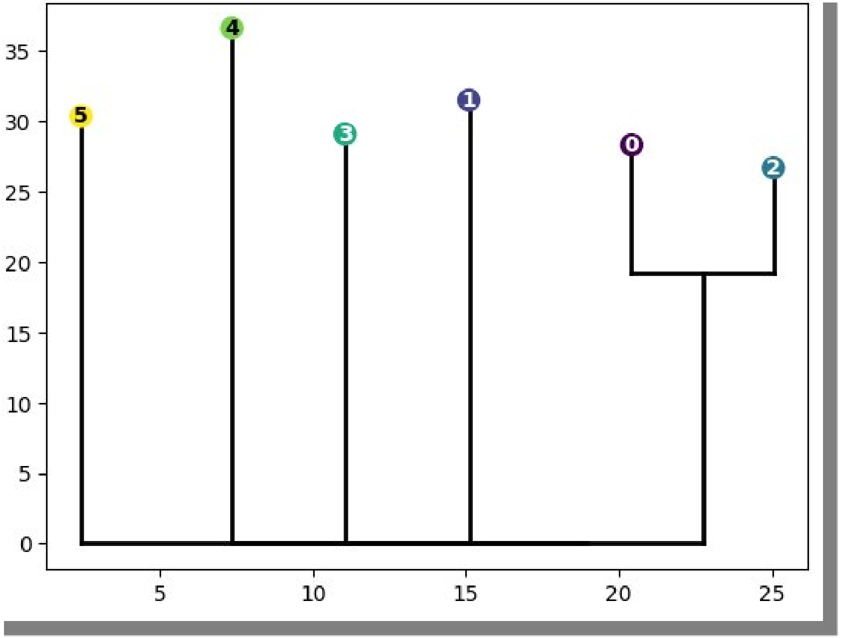
Without feature selection



Cluster to Cancer Mapping:
Cluster -1: (BRCA: 186
KIRC: 35, LUAD: 2)
Cluster 0: (BRCA: 1)
Cluster 1: (LUAD: 20)
Cluster 2: (PRAD: 7)
Cluster 3: (PRAD: 7)
Cluster 4: (KIRC: 110)
Cluster 5: (PRAD: 3)

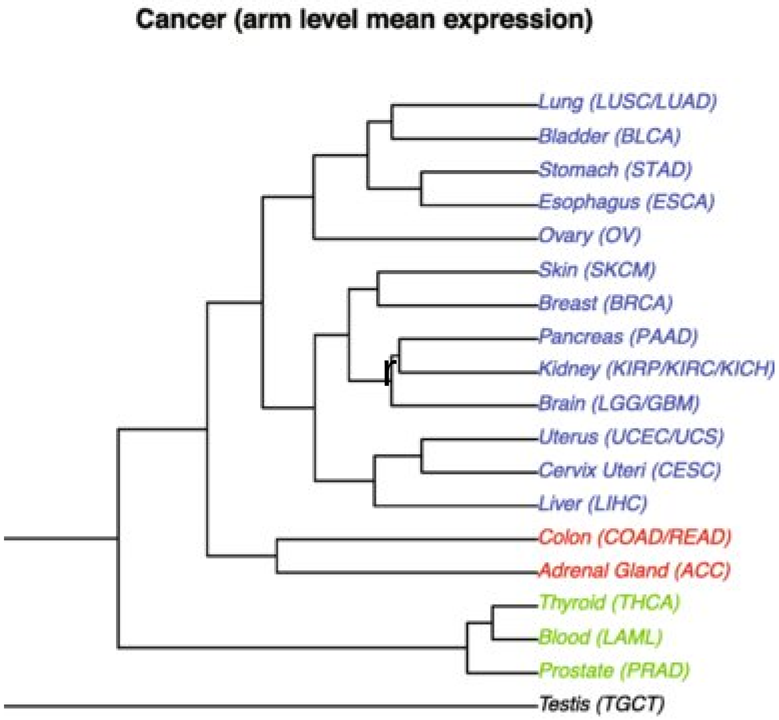
Cluster 6: (LUAD: 118)
Cluster 7: ('COAD:78, LUAD: 1)
Cluster 8: (BRCA: 106)
Cluster 9: (KIRC: 1)
Cluster 10: (PRAD: 119)
Cluster 11: (BRCA: 2)
Cluster 12: (BRCA: 5)

Subset feature selection



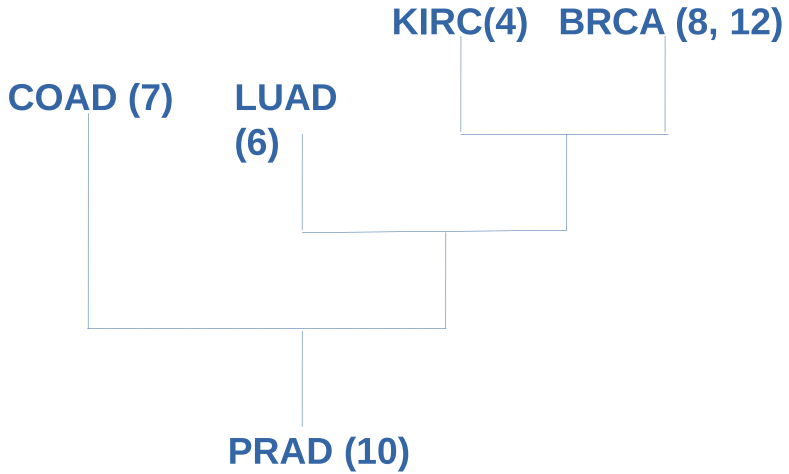
Cluster to Cancer Mapping:
Cluster -1: (LUAD: 63)
Cluster 0: (LUAD: 71)
Cluster 1: (KIRC: 146)
Cluster 2: (LUAD: 7)
Cluster 3: (BRCA: 300)
Cluster 4: (PRAD: 136)
Cluster 5: (COAD:78)

Hierarchical cluster analysis of cancers and normal tissues.



hclust” utility function available in R.
Ref: Patkar et al. Genome Medicine (2021) 13:93

Reducing to the cancer cell types in our data set



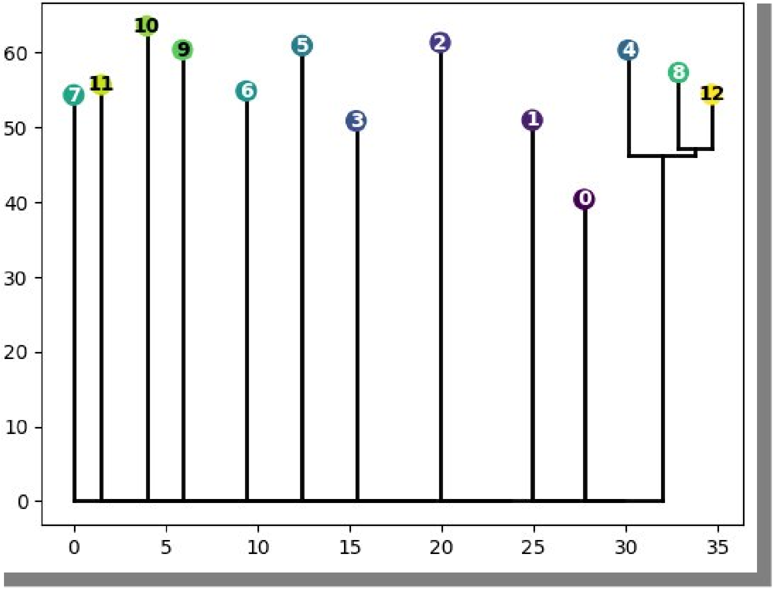
KIRC(4)

BRCA (8, 12)

COAD (7)

LUAD (6)

PRAD (10)



Conclusion

Good Performance on Original Features: The results indicate that DPA performs well even without feature selection, maintaining high ARI and NMI scores. This suggests that DPA can effectively handle high-dimensional data and still produce meaningful clusters.

Stability of Results: The stability of DPA's performance across different feature sets (original and subset) underscores its robustness and reliability..

The low value for value for Silhouette score is due to the fact that by accepting a low level of confidence (low value of Z) we increase the probability of observing some spurious clusters not well separated as it is shown in the clustering results.

DPA outperforms over HDBSCAN (without feature selection) even including Halo points. Although, HDBSCAN is designed to detect clusters with different densities it could be labeling as noise low density packed data points that are part of a cluster. In fact, this can be seen when we compare the clustering plots for the original features and the selected features. We can see that the clusters in the elected features are more compacted or with higher density and the number of points selected as noise are 48 compared with the 535 points labeled as noise for the clustering with original features.

Conclusion

Although, the results of Spectral Clustering were slightly better than DPA for the three metrics, it requires specifying the number of clusters (K), which require additional methods, as eigenvalue gap analysis, to determine the appropriate number of clusters. This dependency can be a limitation in practical applications where such information is not readily available.

Non-parametric Nature: Unlike methods such as k-means or Spectral Clustering, which require specifying the number of clusters beforehand, DPA automatically detects the number of clusters. This makes DPA highly suitable for exploratory data analysis, where the number of clusters is not known in advance.

Lastly, DPA algorithm describes the topography of the data and the hierarchical relation among clusters, which adds a valuable information to the data analysis. As we saw in the dendrogram, DPA was able to “recognize” hierarchical relation among different types of cancer cells. This information can be used in diagnosis or cancer treatments strategies.