Title: Group 1 Case Study 1
Model Selection for Clustering

Name: Pitchakorn Pichetnaruemit
Titichaya Vongbusayamas
Kawin Wachirachaiyakarn
ID: 2848792P
2846125V
2924632w

Introduction

In this case study, we aim to analyse the resulting clusters obtained by utilising various clustering algorithms and parameter configurations. The experiments applied multiple clustering algorithms to the Human Colorectal Cancer (CRC) and normal tissue dataset to identify inherent patterns using a given feature set. Our overarching question is whether these clustering approaches, when applied to distinct feature sets like Pathology-GAN and ResNet50, can reveal inherent patterns in the dataset that may contribute to our understanding of cancer causes.

Therefore, we have two representations of distinct feature sets: Pathology-GAN and ResNet50 in principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) projections. The study employs the K-Means, Hierarchical, and Gaussian Mixture Model clustering algorithms. The evaluation metrics for clustering include silhouette and Davies-Bouldin index as performance measurements for intrinsic measures. These metrics help determine the number of clusters and resolutions for K-Means and Louvain algorithms, respectively. Then, we test our approach by using extrinsic performance measures, such as the rand index and v-measure, to confirm that parameters from previous steps are critical in distinguishing the data into groups, which may lead to causes of cancer. Lastly, we compute the mean of the optimal k from the elbow method and each performance measure. This collective mean is identified as the final optimal k and performs cluster visualisation for each clustering algorithm.

Methods

Feature sets representations

- **Pathology-GAN** in UMAP and PCA
- **ResNet50** in UMAP and PCA

Clustering algorithm and its parameters

- **K-Means:** n_clusters
    - An unsupervised machine learning algorithm for clustering by partitioning datasets into k clusters where each data point joins the cluster with the closest average.
    - **n_clusters:** specifying the number of clusters
- **Hierarchical Clustering (Agglomerative):** n_clusters, metric, linkage

    - An unsupervised machine learning algorithm for clustering that builds a hierarchy of clusters by iteratively merging or splitting them.
    - **n_clusters:** specifying the number of clusters
    - **metric:** defining the metric used to compute the linkage between clusters. In this case study, we use “euclidean” for all experiments.
    - l**inkage:** determining the linkage criterion to decide the merging strategy. In this case study, we use “single” for all experiments.

- **Gaussian Mixture Model:**
    - An unsupervised machine learning clustering algorithm which assumes that the data is generated from a combination of multiple gaussian distributions but different in mean and covariance parameters. Thus, the model combines those individual distributions to clustering the data.
    - **n_components:** specifying the number of clusters

Evaluation Metrics:

- **Elbow method:** determining the most optimal k for each chosen algorithm.
- **Score visualisation:** Scores obtained from each evaluation metric are visualised to assist in determining the most optimal k.
    - **Intrinsic Measures**: do not require ground truth labels
        - **Silhouette coefficient (max):** A measure of how similar an object is to its own cluster compared to other clusters. The maximum value is sought.
        - **Davie Bouldin score (min):** A measure of the average similarity between each cluster and its most similar one. The minimum value is sought.
    - **Extrinsic Measures**: require ground truth labels
        - **Rand index (max):** A measure of the similarity between two data clusterings. The maximum value is sought.
        - **V-Measure (max):** A metric that balances homogeneity and completeness. The maximum value is sought.

Final optimal k:

We gather the optimal k values from each metric, including metrics like the elbow method, silhouette coefficient, Davie Bouldin score, Rand index, and V-Measure. The obtained optimal k values are then summed up with the elbow value. The sum is divided by the total number of optimal k values (in this case, 5 metrics) to calculate the mean of the optimal k. This mean value is subsequently used in the next step for visualisation.

Calculate mean of the optimal k from each evaluation metric

- final_optimal_k = floor(elbow_k + silhouette_k + davie_k + rand_k + v_meansure_k / 5)

Cluster visualisation:

- Visualising clusters using PCA to represent the clustering solutions in 3D space. Each of the final optional k clusters is assigned a seeded distinct random for better interpretation.

Results

The first result of our work is the most optimal k(dashed vertical line) that we got from the elbow method for different clustering algorithms. The majority of optimal results centres around 8, although certain algorithms with some specific dataset indicate a value of k equal 2. Additionally, the four score visualisation graphs also show the same unusual k number cluster which will be discussed later. Then we compute the floored mean of all five k answers from the previous step to visualise the data in k clusters in the graphs below.

We utilise the KElbowVisualizer library to visualise the optimal k value determined by the elbow method. Subsequently, we create a plot featuring four metrics—Silhouette, Davies-Bouldin, Rand index, and V-measure—in individual subplots for straightforward comparison. Each subplot displays its optimal k obtained through score performance evaluation. Finally, we select the optimal k based on the maximum and minimum values, adhering to the method's specific requirements.

K-Means Evaluation and Optimal Parameter

Figure 1-4 illustrates the performance of the KMeans model across varying numbers of clusters, along with the optimal parameter for each metric.

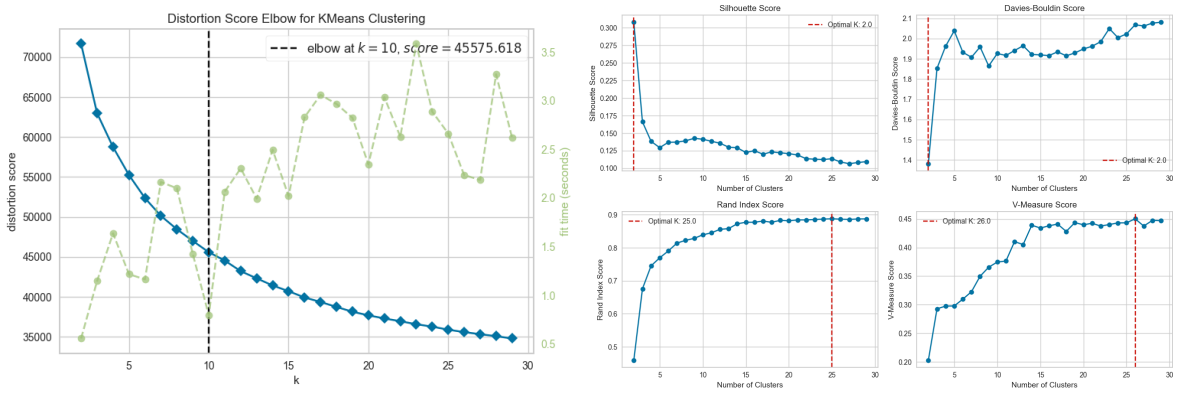

Figure 1. Optimal K from different evaluation metrics of KMeans with PGE PCA features

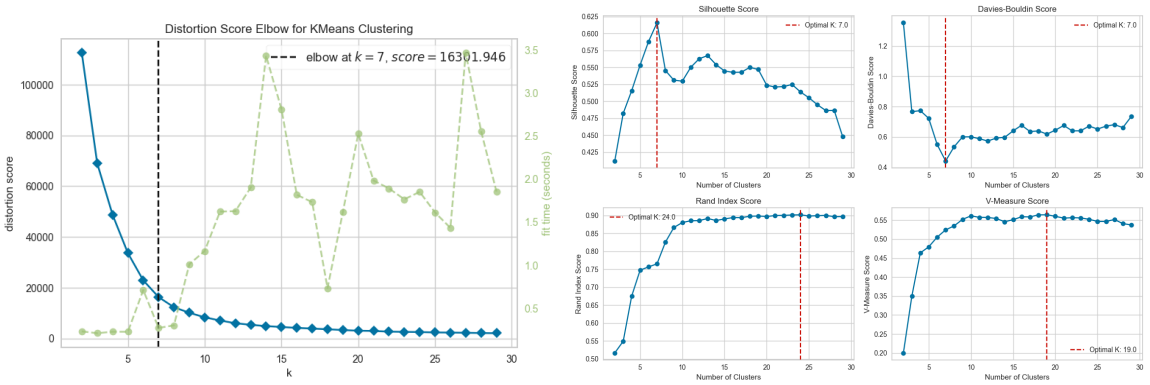

Figure 2. Optimal K from different evaluation metrics of KMeans with PGE UMAP features

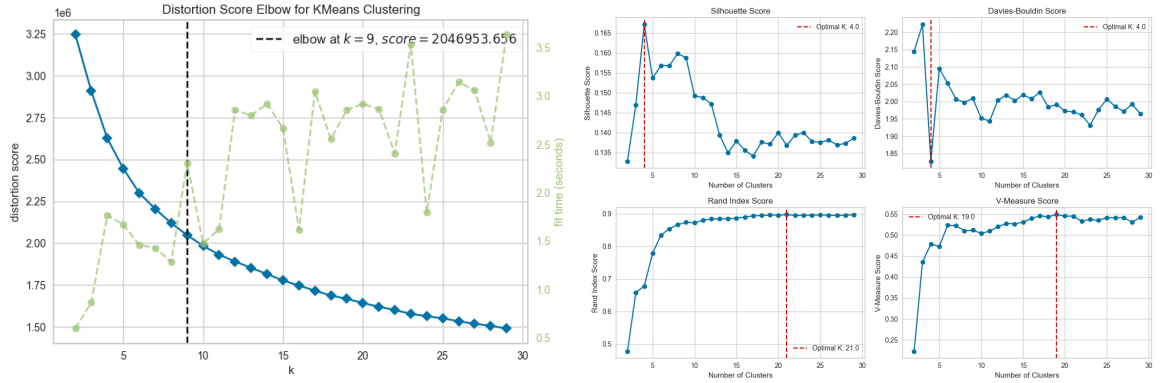

Figure 3. Optimal K from different evaluation metrics of KMeans with ResNet50 PCA features

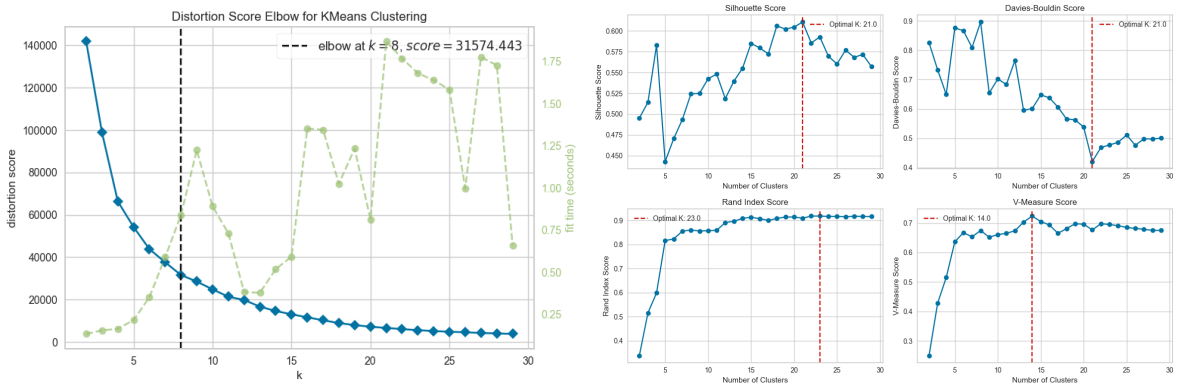

Figure 4. Optimal K from different evaluation metrics of KMeans with ResNet50 UMAP features

Hierarchical Clustering (Agglomerative) Evaluation and Optimal Parameter

Figure 5-8 depicts the performance of Hierarchical Clustering across different numbers of clusters, including the optimal parameter for each metric.

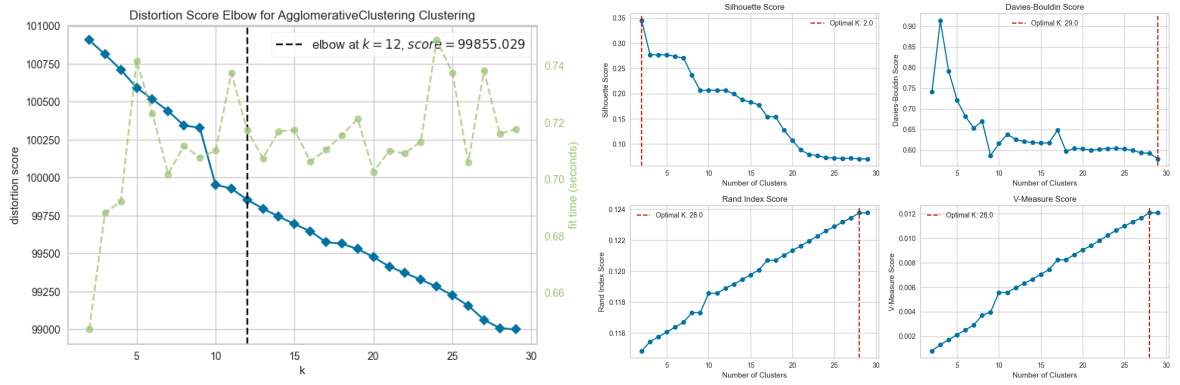

Figure 5. Optimal K from different evaluation metrics of Hierarchical Clustering with PGE PCA features

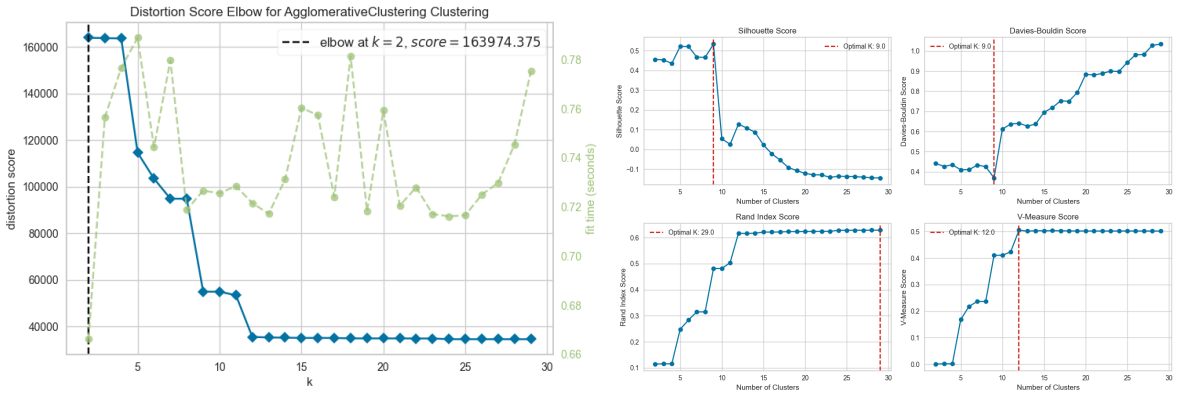

Figure 6. Optimal K from different evaluation metrics of Hierarchical Clustering with PGE UMAP features

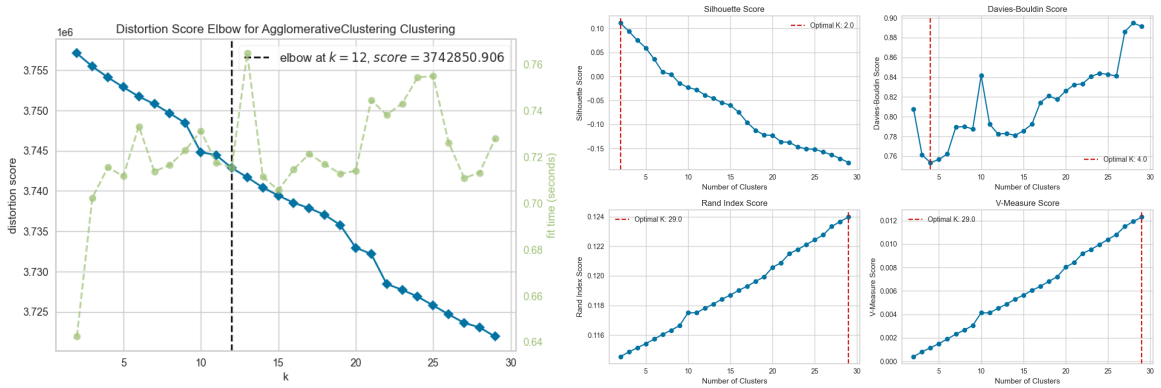

Figure 7. Optimal K from different evaluation metrics of Hierarchical Clustering with ResNet50 PCA features

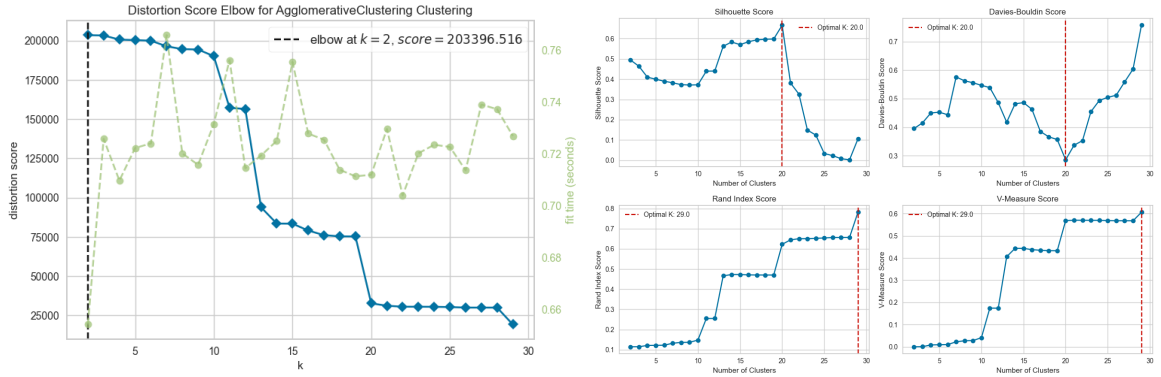

Figure 8. Optimal K from different evaluation metrics of Hierarchical Clustering with ResNet50 UMAP features

Gaussian Mixture Model Evaluation and Optimal Parameter

Figure 9-12 displays the performance of the Gaussian Mixture Model across varying numbers of clusters, along with the optimal parameter for each metric.

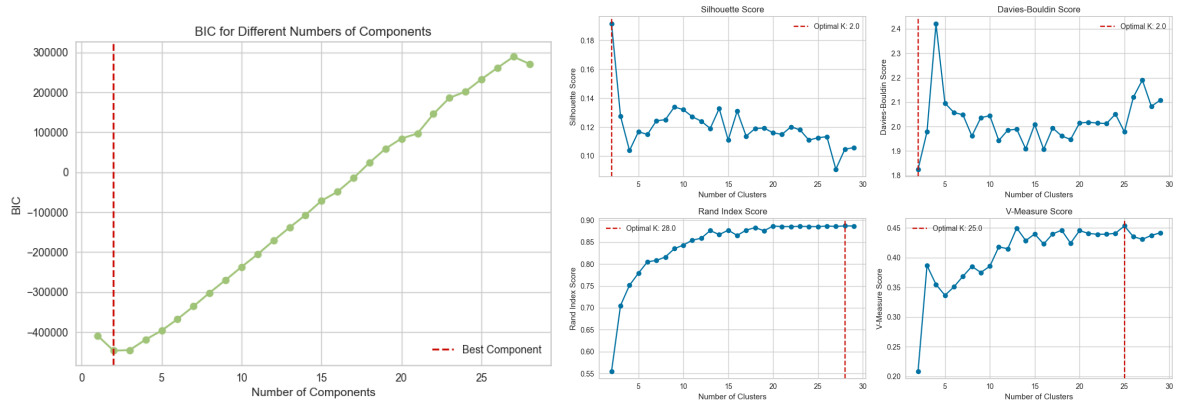

Figure 9. Optimal K from different evaluation metrics of GMM with PGE PCA features

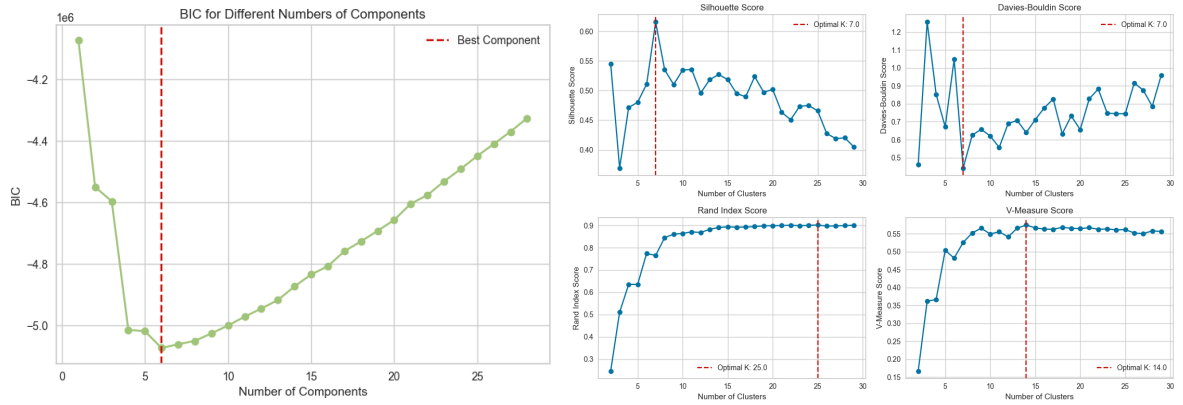

Figure 10. Optimal K from different evaluation metrics of GMM with PGE UMAP features

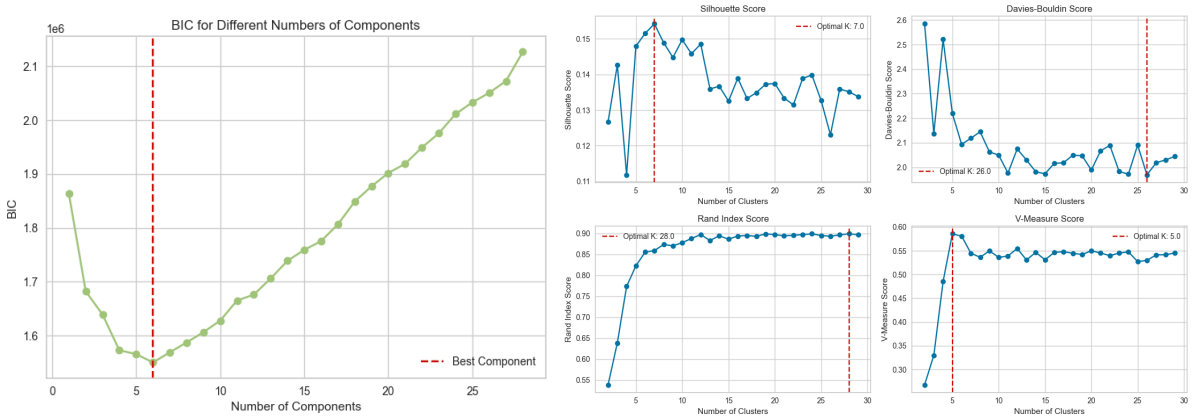

Figure 11. Optimal K from different evaluation metrics of GMM with ResNet50 PCA features

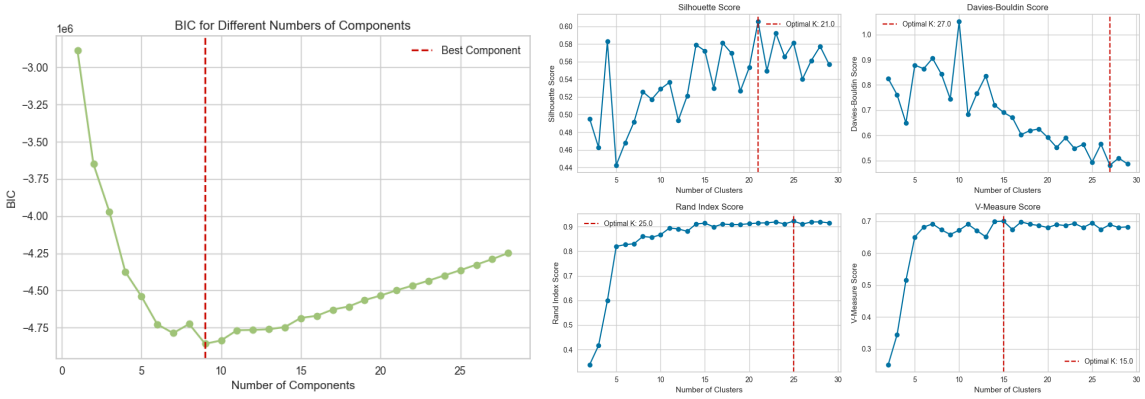

Figure 12. Optimal K from different evaluation metrics of GMM with ResNet50 UMAP features

K-Means Result Clusters

Figure 13-14 displays the clusters created by the KMeans model with the optimal number of clusters using both PCA and UMAP from PGE and Resnet50 respectively.

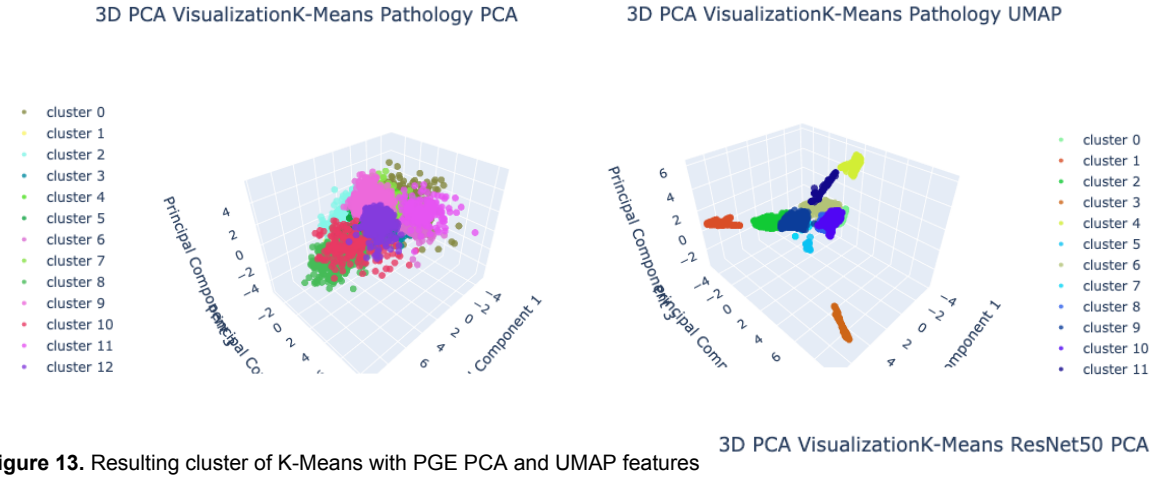

Figure 13. Resulting cluster of K-Means with PGE PCA and UMAP features

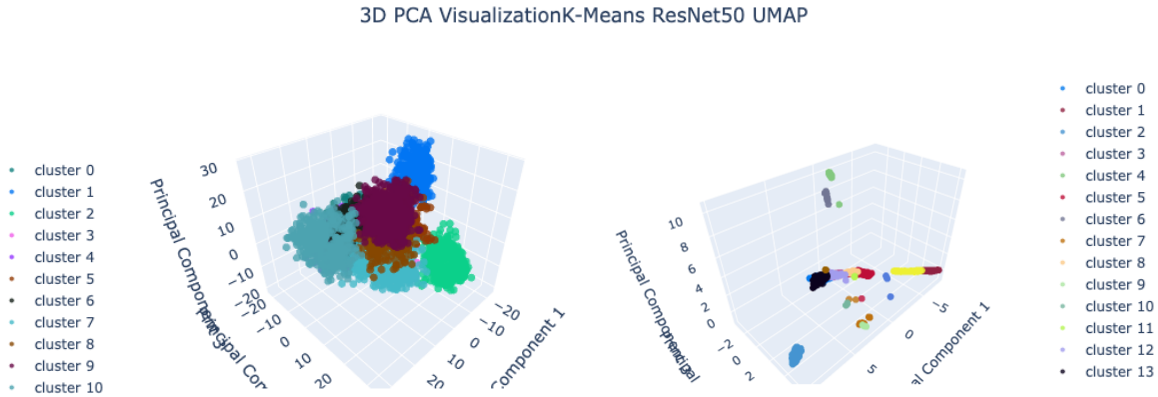

Figure 14. Resulting cluster of K-Means with ResNet50 PCA and UMAP features

Hierarchical Clustering Result Clusters

Figure 15-16 displays the clusters created by the Hierarchical Clustering model with the optimal number of clusters using both PCA and UMAP from PGE and Resnet50 respectively.

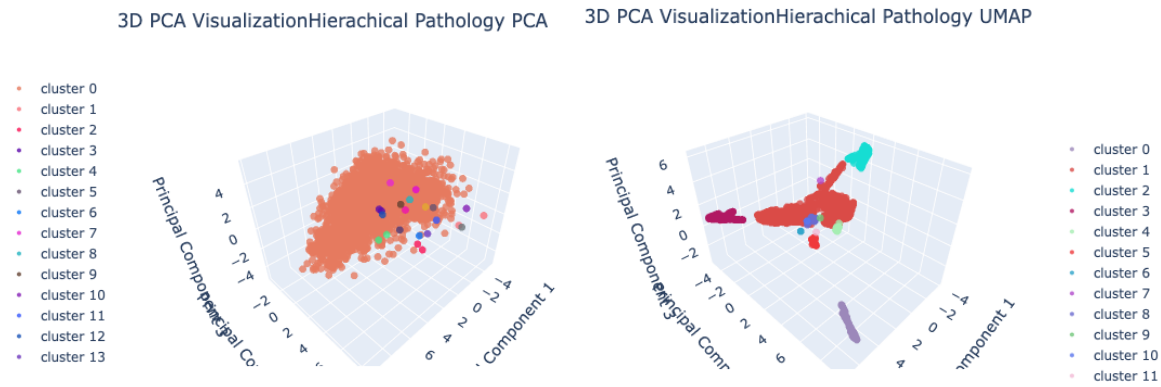

Figure 15. Resulting cluster of Hierarchical Clustering with PGE PCA and UMAP features

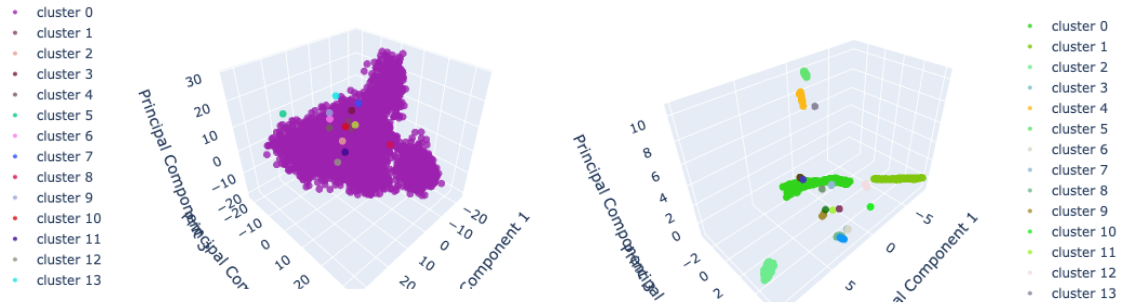

Figure 16. Resulting cluster of Hierarchical Clustering with ResNet50 PCA and UMAP features

Gaussian Mixture Model Result Clusters

Figure 17-18 displays the clusters created by the GMM model with the optimal number of clusters using both PCA and UMAP from PGE and Resnet50 respectively.

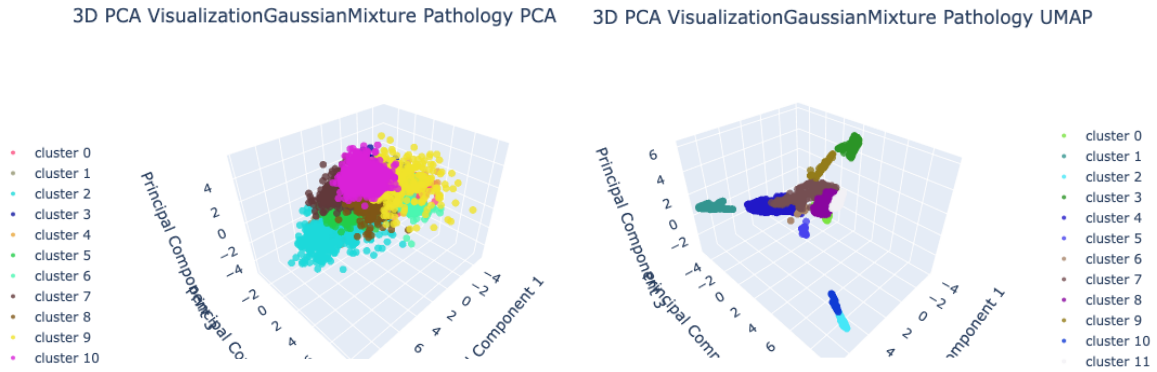

Figure 17. Resulting cluster of GMM with PGE UMAP features

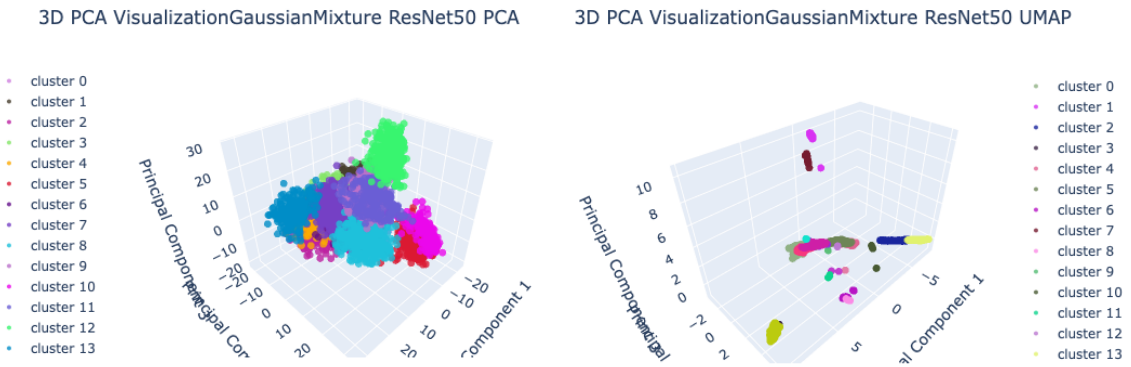

Figure 18. Resulting cluster of GMM with ResNet50 PCA and UMAP features

Discussion

The results suggest that the performance of clustering algorithms varies depending on the feature set used, such as Pathology-GAN and ResNet50. The hypothesis that some algorithms struggle with outlier data, potentially treating each outlier as a separate cluster, aligns with the observation that certain algorithms perform poorly when faced with highly concentrated data points in a single cluster.

Specifically, the conclusion drawn from the comparison is that Hierarchical clustering, especially when utilising PCA features, performs inadequately as it tends to concentrate all data points in a single cluster. On the other hand, K-Means and Gaussian Mixture Model

exhibit more balanced cluster division. This discrepancy in performance underscores the sensitivity of clustering algorithms to the nature of the data and the importance of choosing an algorithm that aligns with the characteristics of the dataset.

To conclude, the varying results across algorithms and feature sets emphasise the need for a nuanced understanding of the data and the algorithm's capabilities for meaningful insights into cancer causes.