Machine Learning& AI Case Study 1

Siyu Na, 2931267N

Xulong Yang, 2870692Y

Xinghan Cheng 2938618C

School of Computing Science

Sir Alwyn Williams Building

University of Glasgow

G12 8RZ

A dissertation presented in part fulfillment of the requirements of the Degree of Master of Science at the University of Glasgow

29 November 2023

1 Background

Hematoxylin and Erosin (H&E) staining is a widely used technique in histology for visualizing cellular structures and tissue morphology. [1] In the context of human colorectal cancer (CRC) and normal tissue, H&E staining is employed to examine and distinguish between healthy and cancerous tissues at a microscopic level.[2]

When a tissue sample from the colorectal region undergoes H&E staining, the hematoxylin dye stains cell nuclei blue-purple, providing contrast and highlighting their presence. On the other hand, eosin stains various cellular components, including cytoplasm and extracellular matrix, in shades of pink or red, allowing differentiation of different tissue structures.

By utilizing H&E staining on histological samples obtained from human colorectal tissues, pathologists and researchers can observe the cellular architecture, cellular arrangements, and overall tissue morphology.[3] This staining technique aids in identifying potential abnormalities, such as tumor formations, alterations in cell structures, and the presence of cancerous cells, within the colorectal tissue samples. The visual information obtained from these stained images is crucial for understanding the histopathology and diagnosing various colorectal conditions, including cancer, based on the observed cellular and tissue characteristics.

2 Introduction

Clustering, being an unsupervised learning technique, relies solely on the inherent structure of the data within a given feature set. Unlike supervised learning, where models learn from labeled data, clustering extracts patterns and groupings from unlabeled data, aiming to uncover inherent relationships or clusters that may exist within the dataset.

In this scenario, the challenge lies in assessing the performance of clustering models without predefined classes or labels. The quality of clusters formed and their sensitivity to various clustering algorithms and feature spaces becomes a complex task.

In this case, we use four data representations, including PathologyGAN, ResNet50, InceptionV3, and VGG16. For each representation, we test both PCA and UMAP projections. We use four clustering algorithms and also a range of the potential parameter(s) that affect the number of clusters.

However, gauging the "quality" of these clusters in an unsupervised setting poses a significant challenge. Metrics like Silhouette Scores and V-measure Scores are utilized to quantify the goodness and consistency of clusters. Silhouette Scores indicate how well-defined the clusters are [4], while V-measure Scores assess the homogeneity

and completeness of the clustering compared to the available representations or ground truth [5].

The comprehensive evaluation involves an intricate exploration of these multiple dimensions: different feature representations, dimensionality reduction techniques, diverse clustering algorithms, and their corresponding parameter variations. By meticulously analyzing the results across these multifaceted approaches, insights into the inherent structures within the data can be extracted, aiding in a deeper understanding of the dataset's underlying patterns and relationships.

3 Methods

3.1 UMAP and PCA：

PCA is a statistical method that transforms a set of possibly correlated variables (multiple indicators of the observed values) into a set of linearly uncorrelated variables through an orthogonal transformation; these new variables are called principal components. [6] It is used to reduce the data set of dimensions at the same time keep a method of original data variance as soon as possible. PCA is usually used to deal with linear relationships and works best when the linear assumptions of the original data are true.

UMAP is a relatively new machine-learning algorithm for health and data exploration. With PCA, UMAP is based on manifold learning, it tries to keep in low dimensional space of high-dimensional data structure of local and global. UMAP is particularly effective for nonlinear data structures, can capture more complex patterns, and excels in visualization.

3.2 Clustering Methods：

3.2.1 K-means

K-means is a simple clustering algorithm. It is the grouping of data points into K clusters, each of which is defined by its centroid (i.e., cluster center). It first assigns each data point to the nearest center of mass.[7] It then iterates by calculating the new centroid of each cluster (the mean of all points in the cluster). This algorithm is fast and suitable for large data sets, but the number of clusters K needs to be specified in advance, and the choice of initial centroid is more critical to the algorithm.

3.2.2 Gaussian Mixture Model

Gaussian mixture model is a probability-based clustering technique which assumes that the data is mixed by several Gaussian distributions. Compared to K-means, GMM is more flexible because it takes into account not only the center of mass, but also the shape and orientation of the data point distribution. In GMM, each cluster is defined by a Gaussian distribution whose parameters are the mean (cluster center) and covariance (describing the shape of the cluster). GMM uses the expectation maximization (EM) algorithm to estimate these parameters and assigns a cluster membership probability to each data point. GMM is able to recognize more complex clustering shapes, but it is computationally expensive. Similar to the K-means algorithm, the number of clusters needs to be specified in advance.

3.2.3 Hierarchical Clustering

Unlike the previous two methods, hierarchical clustering is an algorithm that does not need to specify the number of clusters in advance. It builds a hierarchy of clusters, which can be either bottom-up aggregation (condensation method) or top-down splitting (splitting method). The algorithm initially treats each data point as a separate cluster, then progressively merges the most similar clusters, then progressively splits the least similar subsets.

3.2.4 Louvain Clustering

Louvain clustering is mainly used for network graph data and is an algorithm based on modularity optimization aimed at discovering high-density subgraphs in graphs. When iterating again, the algorithm first finds the best community assignment for each node, then merges the currently found community into the new node, and repeats the process. It is able to quickly detect community structure in large networks and does not require the number of communities to be specified in advance. Louvain's algorithm is particularly suitable for network structure data, such as social networks or biological networks, but is not commonly used for general clustering problems.

3.3 Validation Methods：

In the evaluation process, the Silhouette Score[8] and the V measure score[9] were used to evaluate the algorithm. Where the Silhouette Score is the ratio of the difference between the degree of separation and the degree of l cohesion to the greater value of the two, that is, the ratio of the average distance between other points in the same cluster and the average distance z between it and all points in the nearest cluster. The result closer to 1 indicates that the algorithm can match its own cluster better. V measure score represents the integrity of fractional atmospheric homogeneity. Homogeneity means that each cluster contains only a single category of data points, and integrity means that all data points in the same category are assigned to the same cluster. A result closer to 1 indicates that the results of the cluster are more consistent with the real label.

4 Design:

For K-means, Gaussian Mixture Model and Hierarchical Clustering, three algorithm models of Scikit-Learn library were used to complete the code. After completing Louvain Clustering, a graph for Louvain Clustering algorithm was constructed by calling networkx library, and then adjust the number of clusters and add edges to the image by changing the threshold of the distance matrix. The community Louvain and Best partition method were used to apply the Louvain community detection algorithm to the constructed graph and determine the number of clusters by counting the number of unique values in the dictionary.

5 Result

The table shows that in almost every dataset, K-means generally has good silhouette scores and V-measure scores. The performance of GMM fluctuates under different circumstances. The silhouette scores and V-measure scores of Agglomerative clustering are close to those of K-means in some cases, but lower in others. Meanwhile, Louvain's scores are generally lower among all the algorithms.

	pca				umap				
PathologyGAN	Metrics	Kmeans	gmm	agglomeerative	louvain	Metrics	Kmeans	gmm	agglomeerative	louvain
	silhouette	0.140692	0.133565	0.123276	0.143758	silhouette	0.536229	0.439436	0.470999	0.398944
	V-measure	0.341082	0.310122	0.423920	0.376787	V-measure	0.514082	0.388531	0.483977	0.399438
ResNet50	Metrics	Kmeans	gmm	agglomeerative	louvain	Metrics	Kmeans	gmm	agglomeerative	louvain
	silhouette	0.163293	0.164692	0.149230	0.162418	silhouette	0.612741	0.490884	0.612741	0.487456
	V-measure	0.558838	0.563265	0.538328	0.516760	V-measure	0.584410	0.526145	0.584410	0.534946
InceptionV3	Metrics	Kmeans	gmm	agglomeerative	louvain	Metrics	Kmeans	gmm	agglomeerative	louvain
	silhouette	0.264440	0.252680	0.230307	0.242540	silhouette	0.510769	0.422578	0.508401	0.527992
	V-measure	0.354245	0.359575	0.424255	0.369732	V-measure	0.414108	0.364431	0.407870	0.349597
VGG16	Metrics	Kmeans	gmm	agglomeerative	louvain	Metrics	Kmeans	gmm	agglomeerative	louvain
	silhouette	0.136721	0.107240	0.122034	0.110536	silhouette	0.510769	0.422578	0.508401	0.527992
	V-measure	0.561777	0.476188	0.583274	0.520908	V-measure	0.414108	0.364431	0.407870	0.349597

Tabel 5.1 Results of all experiments

For the PathologyGAN dataset, the UMAP dimensionality reduction method yields better results in terms of silhouette scores and V-measure scores, especially with the K-means method. This may be due to the PathologyGAN data having more pronounced non-linear characteristics. The visualization image is shown as follows:

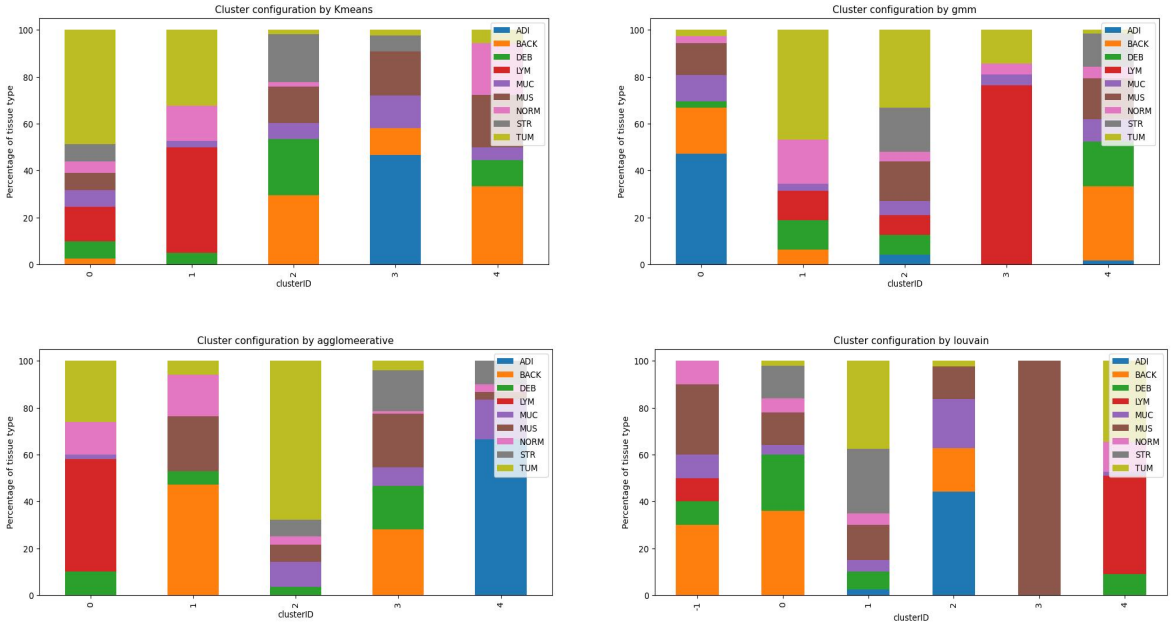

Figure 5.2 Visulisation of clustering methods with PCA

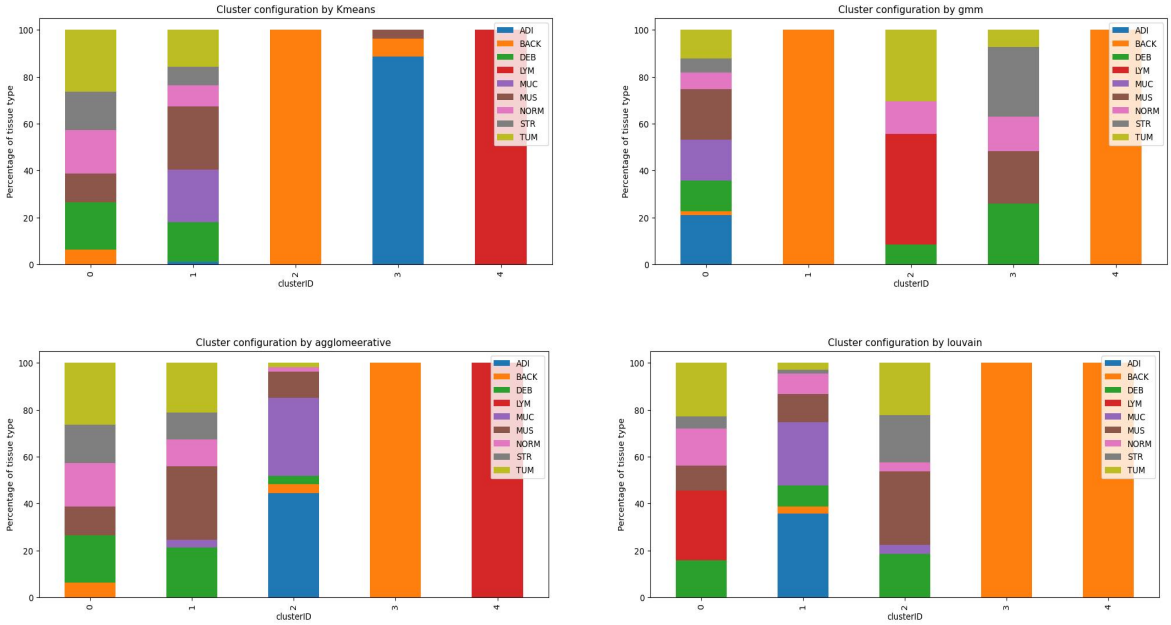

Figure 5.3 Visulisation of clustering methods with UMAP

For the ResNet50 dataset, the UMAP dimensionality reduction method also performs better. In this dataset, the silhouette scores and V-measure scores of Agglomerative clustering are close to those of the K-means method. This indicates that in this case, Agglomerative clustering has performance comparable to K-means. The visualization

image is as follows:

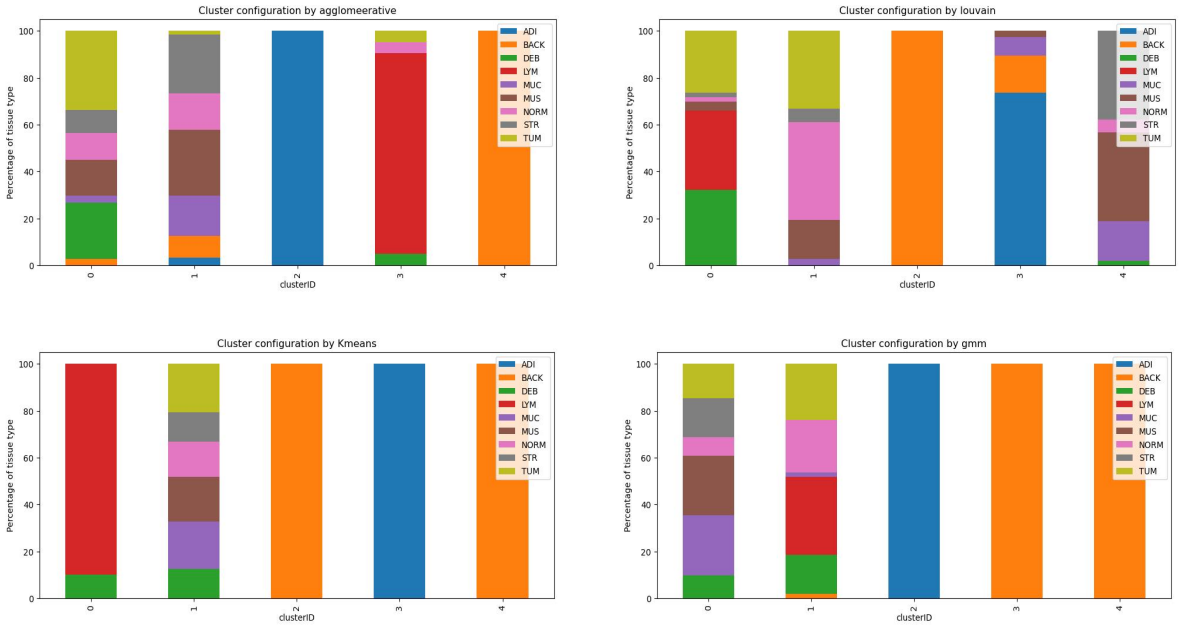

Figure 5.4 Visulisation of clustering methods with PCA

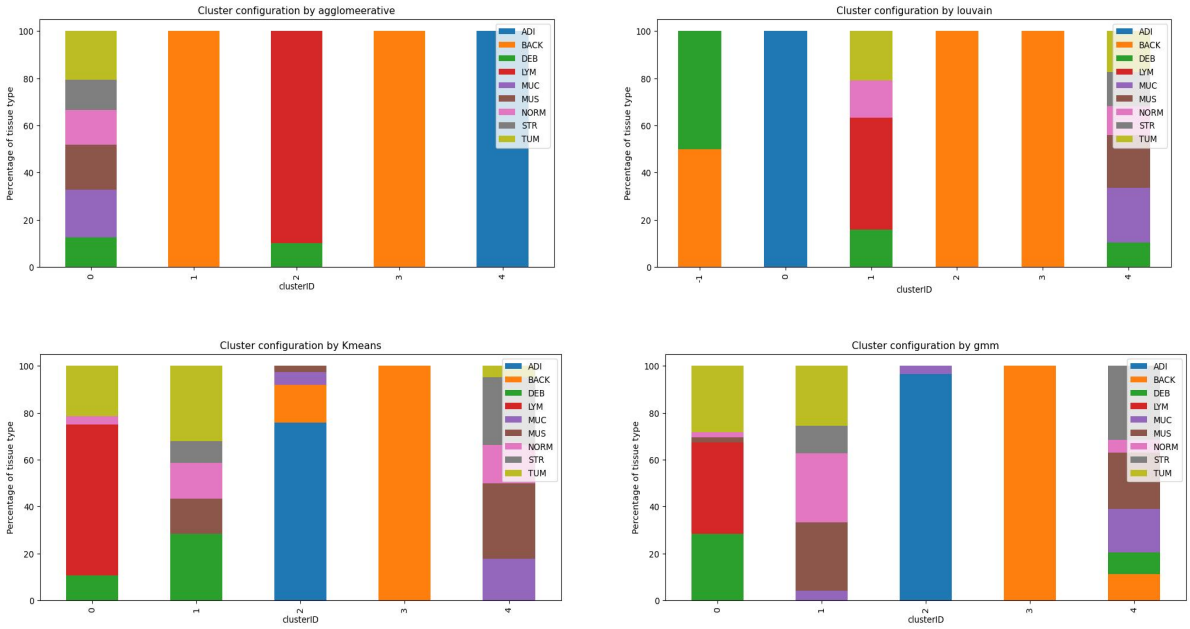

Figure 5.5 Visulisation of clustering methods with UMAP

For the InceptionV3 dataset, using the UMAP dimensionality reduction method continues to yield better results. Notably, the Louvain clustering method achieves higher silhouette scores in this dataset, but its V-measure scores remain relatively low. The visualization image is as follows:

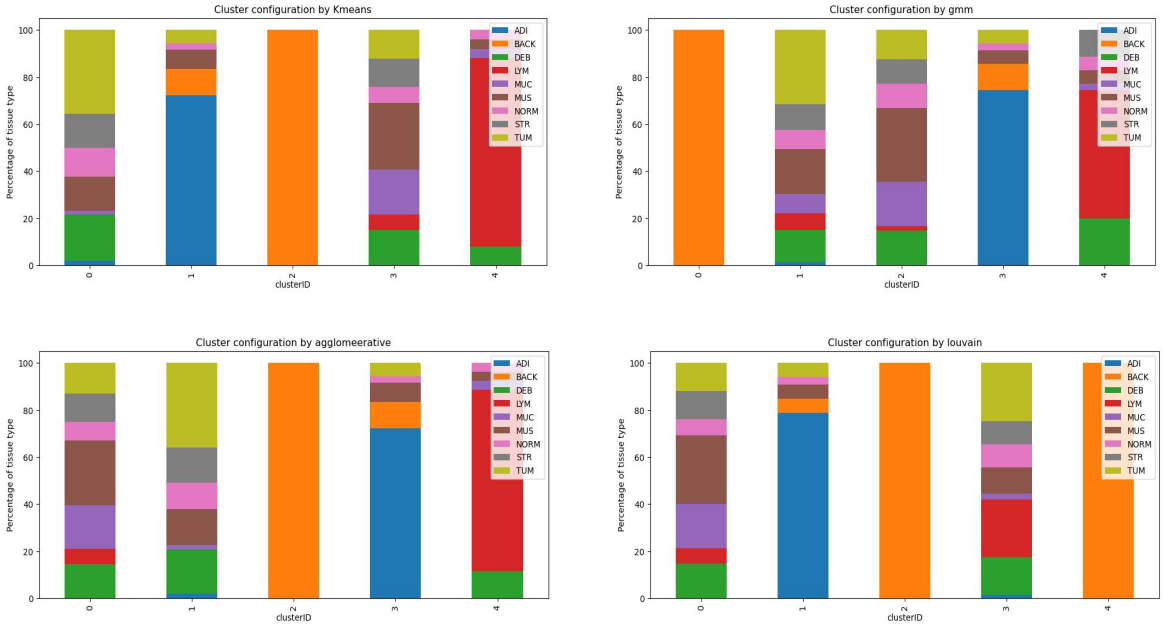

Figure 5.6 Visulisation of clustering methods with PCA

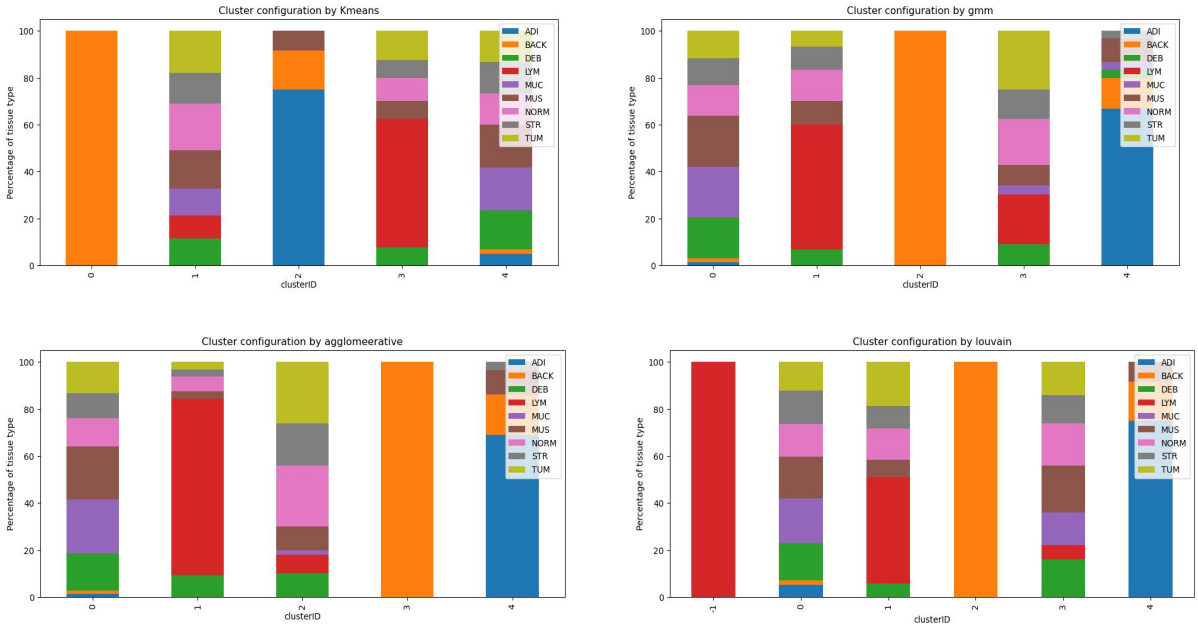

Figure 5.7 Visulisation of clustering methods with UMAP

For the VGG16 dataset, using UMAP for dimensionality reduction significantly improves the silhouette scores of all four algorithms. After its application, there is a notable increase in the silhouette scores for the Louvain algorithm. However, the use of UMAP results in a decrease in the V-measure scores for all four algorithms. The image visualization is as follows:

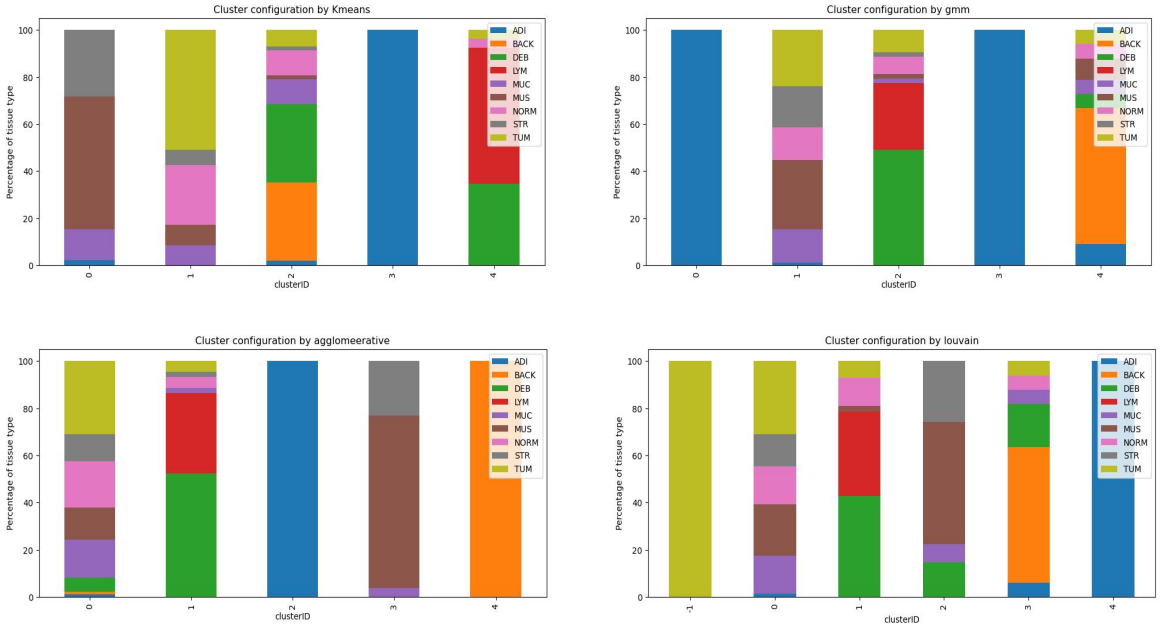

Figure 5.8 Visulisation of clustering methods with PCA

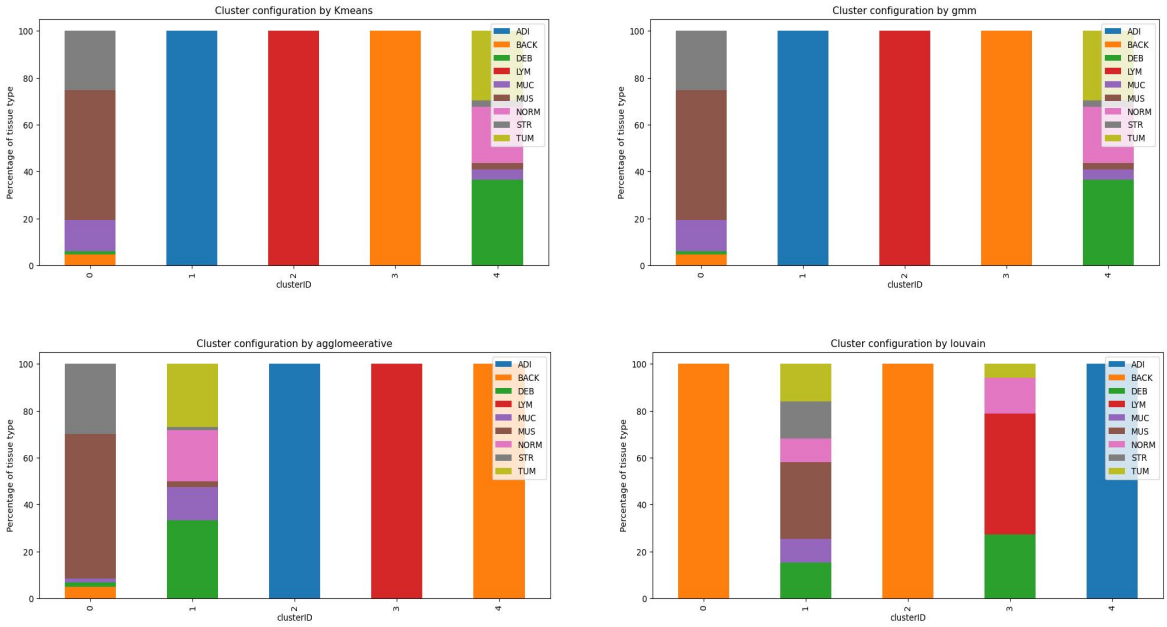

Figure 5.9 Visulisation of clustering methods with UMAP

6 Discussion

The performance of the four algorithms across the four datasets and two dimensionality reduction methods does not show a clear superiority. Among them, the K-means method is the most stable, possibly because it is better suited to the structure of these datasets. On the other hand, the Louvain clustering algorithm generally performs poorly across these datasets, likely because it is primarily designed for community detection and may not be suitable for traditional clustering tasks, especially on non-network data.

Additionally, using the UMAP dimensionality reduction method significantly improves the silhouette scores. This could be due to UMAP's ability to preserve both local and global structures of the data, making clusters more distinct and evident, thereby improving silhouette scores. UMAP might also intensify the inherent clustering tendencies in the data, leading to more compact clusters with reduced internal distances and increased distances between clusters, thus enhancing silhouette scores.

However, in the VGG16 dataset, the V-measure scores of all four algorithms decrease after using the UMAP method.[10] This might be due to UMAP altering some of the original category relationships while preserving the data's local structure. The dimensionality reduction process might introduce noise or lose features crucial for clustering, leading to clusters inconsistent with the actual categories. To improve V-measure scores, further refinement of the UMAP method might be necessary to better match the characteristics of the VGG16 dataset.

Reference

[1] Pote, A., Boghenco, O. & Marques-Ramos, A. Molecular analysis of H&E- and Papanicolau-stained samples—systematic review. Histochem Cell Biol 154, 7–20 (2020).

[2] Mokhtari Z, Amjadi E, Bolhasani H, et al. CRC-ICM: Colorectal Cancer Immune Cell Markers Pattern Dataset[J]. arXiv preprint arXiv:2308.10033, 2023.

[3] Pote A, Boghenco O, Marques-Ramos A. Molecular analysis of H&E-and Papanicolau-stained samples—systematic review[J]. Histochemistry and Cell Biology, 2020, 154: 7-20.

[4] de Moraes, Wilson M A M et al, "Carbohydrate Loading Practice in Bodybuilders: Effects on Muscle Thickness, Photo Silhouette Scores, Mood States and Gastrointestinal Symptoms," Journal of Sports Science & Medicine, vol. 18, (4), pp. 772-779, 2019.

[5] T. H. Nguyen et al, "CLUSTERING VIETNAMESE CONVERSATIONS FROM FACEBOOK PAGE TO BUILD TRAINING DATASET FOR CHATBOT," Jordanian Journal of Computers and Information Technology, vol. 8, (1), pp. 1-17, 2022.

[6] Shlens J. A tutorial on principal component analysis[J]. arXiv preprint arXiv:1404.1100, 2014.

[7] Jin, X., Han, J. (2011). K-Means Clustering. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.

[8] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, 2020, pp. 747-748

[9] Rosenberg A, Hirschberg J. V-measure: A conditional entropy-based external cluster evaluation measure[C]//Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). 2007: 410-420.

[10] Ahmed M, Seraj R, Islam S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.