Title: Case study 1 Model selection for clustering

Name: Akash Ananda Kumar Murali ID: 2768146M

Name: Hao Chen ID: 2691051C

Name: Jiachen Dong ID: 2789896D

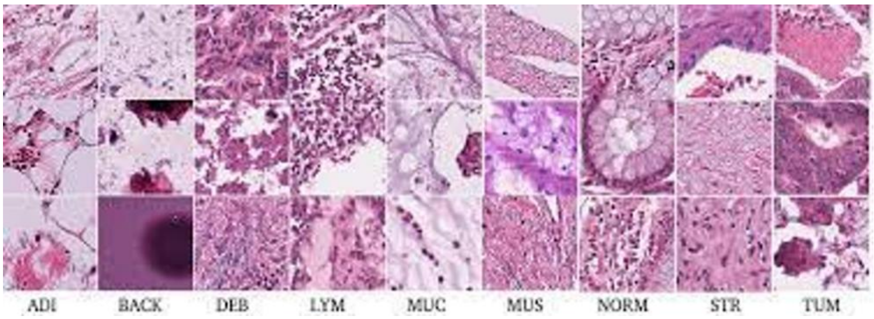
Name: Wenbin Shang ID: 2743882S

Name: Xinyu Guan ID: 2764666G

1.Introduction

1.1 background and data

In this case we focus on cluster analysis of the data set. The data set here is Removed tissue analysed under digital microscope obtained from Cancer diagnosed by biopsy，also the image size is 20Gb on average ~100000x100000 pixels. So in this Case we mainly process the images by first breaking whole slide images (WSIs) into small patches and then performing a series of data processing, such as dimensionality reduction or feature extraction and finally clustering gives statistical summary of visual features.The dataset is 5,000 colorectal cancer tissue patches and includes 9 tissue types, including Adipose (ADI),background (BACK),debris (DEB),lymphocytes (LYM),mucus (MUC),smooth muscle (MUS),normal colon mucosa (NORM),cancer-associated stroma (STR),colorectal adenocarcinoma epithelium (TUM).



In the provided data, Feature extraction and preprocessing has already been done. Firstly use PathologyGAN or ResNet50/InceptionV3/VGG16 to process the complex images and then use PCA and UMAP to simplify it, so basically we get four 2x5000 of 100-d vectors data set.

1.2 clustering

Our main task in this case is to Choosing the best model candidate, and the reason

Why do we need model selection for clustering is that we need to Group similar objects together,make some constraints on clusters and understand the structure of a dataset.

As for How to select the best model for clustering, we will consider lots of factor, for instance,define objective (e.g. accuracy, minimise false positives, etc),complexity, computability, ease of Implementation and so on.

2.Methodology

This section introduces the two classification algorithms we use, describes the content and steps of each algorithm, and describes the role of important parameters in the code.
Clustering is the process of classifying and organising members of a data set that are similar in some way. Clustering is a technique for discovering this intrinsic structure, and clustering techniques are often referred to as unsupervised learning.

2.1 K-means
K-means is our most common Euclidean distance-based clustering algorithm, which assumes that the closer two targets are to each other, the greater the similarity. K-means clustering is the best known algorithm for partitioning clusters, and its simplicity and efficiency make it the most widely used of all clustering algorithms. Given a collection of data points and a desired number of clusters k, with k specified by the user, the k-mean algorithm iteratively partitions the data into k clusters based on some distance function.

There are some advantages to use K-means:
a. Easy to understand, clustering works well, although it is locally optimal, but often a local optimum is sufficient and
b. the algorithm ensures good scalability when dealing with large data sets.
c. works very well when the clusters are approximately Gaussian distributed.
d. The complexity of the algorithm is low.

There are also some disadvantages:
a. K values need to be set artificially and different K values yield different results.
b. Sensitivity to the initial cluster centre, with different results obtained for different selections.
c. Sensitivity to outliers.
d. samples can only be grouped into one category and are not suitable for multi-classification tasks.
e. unsuitable for classifications that are too discrete, classifications with unbalanced sample classes, and classifications of non-convex shapes.

2.1.1 Algorithm steps

So the steps of the K-means algorithm are as follow:

Select the initial k samples as the initial clustering centres;

For each sample in the dataset, calculate its distance to the k cluster centres and assign it to the class corresponding to the cluster centre with the smallest distance.

For each class, recalculate its cluster centre (i.e. the centre of mass of all samples belonging to that class).

Repeat the above 2 3 steps until some stopping condition (number of iterations, minimum error change, etc.) is reached.

2.1.2 Parameter

In the code implementation of the clustering algorithm, the most important parameter is the k value, and we can change the number of categories by changing k.This is because once we have defined k, the algorithm randomly selects K objects as the initial clustering centres. The distance between each object and each seed cluster centre is then calculated and each object is assigned to the cluster centre nearest to it. The clustering centres and the objects assigned to them then represent a cluster. Once all the objects have been assigned, the cluster centres for each cluster are recalculated based on the existing objects in the cluster. This process is repeated until a termination condition is met.

2.2 Louvain Community Detection

The Louvain algorithm is a modularity-based community discovery algorithm. The basic idea is that nodes in the network try to traverse the community labels of all their neighbours and select the community label that maximises the modularity increment. After maximising the modularity, each community is seen as a new node and repeated until the modularity is no longer increasing.

There are some advantages for using Louvain:

a. Low time complexity, suitable for large-scale networks.

b. Stable community segmentation results, with specific metrics to be able to evaluate good or bad community segmentation.

c. The limitation of modularity resolution is eliminated. Since modularity is a global metric, it is difficult to find small communities when maximising, and it is easy to merge small communities. The first iteration of the algorithm, however, uses a single node as the granularity of the community, circumventing this problem.

d. It naturally comes with hierarchical community division results, and the community division results of each iteration can be retained as an intermediate result of community division, which is optional.

There are also some disadvantages, like,The community is too large to converge in time. If we compare the modularity to a loss function, Fast Unfolding's greedy approach in the modularity optimisation phase can easily "overfit" the entire community. Because Fast Unfolding is a point traversal, it is easy to add some peripheral points to the otherwise compact community, leading to some incorrect merging. This partitioning is sometimes good in a local view, but bad in a global view.

2.2.1 Algorithm steps

1, initially treat each vertex as a community, the number of communities is the same as the number of vertices.
2. Merge each vertex with its neighboring vertices in turn, and calculate whether their maximum modularity gain is greater than 0. If it is greater than 0, put the node into the community of the neighboring node with the largest modularity gain.
3. Iterate the second step until the algorithm is stable, i.e. the communities to which all vertices belong no longer change.
4. Compress all nodes of each community into one node, the weight of the points within the community is transformed into the weight of the new node ring, and the weight between communities is transformed into the weight of the new node edge.
5. Repeat steps 1-3 until the algorithm is stable.
Louvain algorithm is very similar to the FN algorithm, I think the biggest difference is that the Louvain algorithm looks at the whole community as a new supernode during the coalescence phase, while the FN algorithm takes the community's point of view to coalesce.

2.2.2 Parameter

The most important parameter in the implementation of this algorithm is the modularity, which is presented in the code as resolution. Modularity is a metric for assessing how well a community network is divided, and it has the physical meaning that the number of contiguous edges of nodes in a community differs only from the number of edges in a random case.

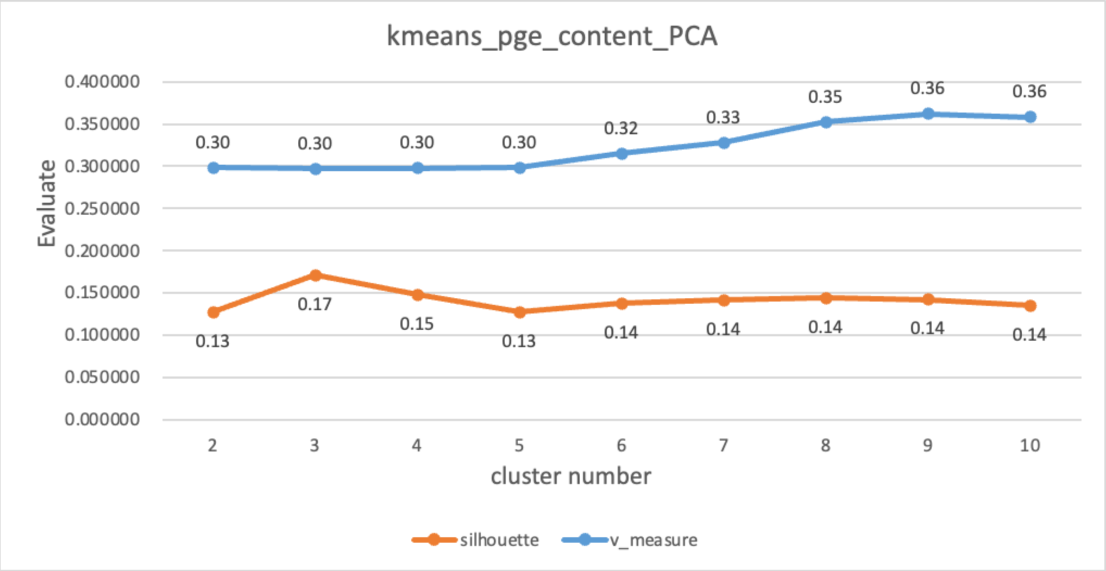
3.Experimental framework

In this section, we describe the process of modelling different algorithms by varying the parameters and evaluating the different models so that the optimal one can be chosen. The analysis was carried out successively on the basis of two different datasets, for each of which we tested the data obtained by PCA and UMAP dimensionality reduction.

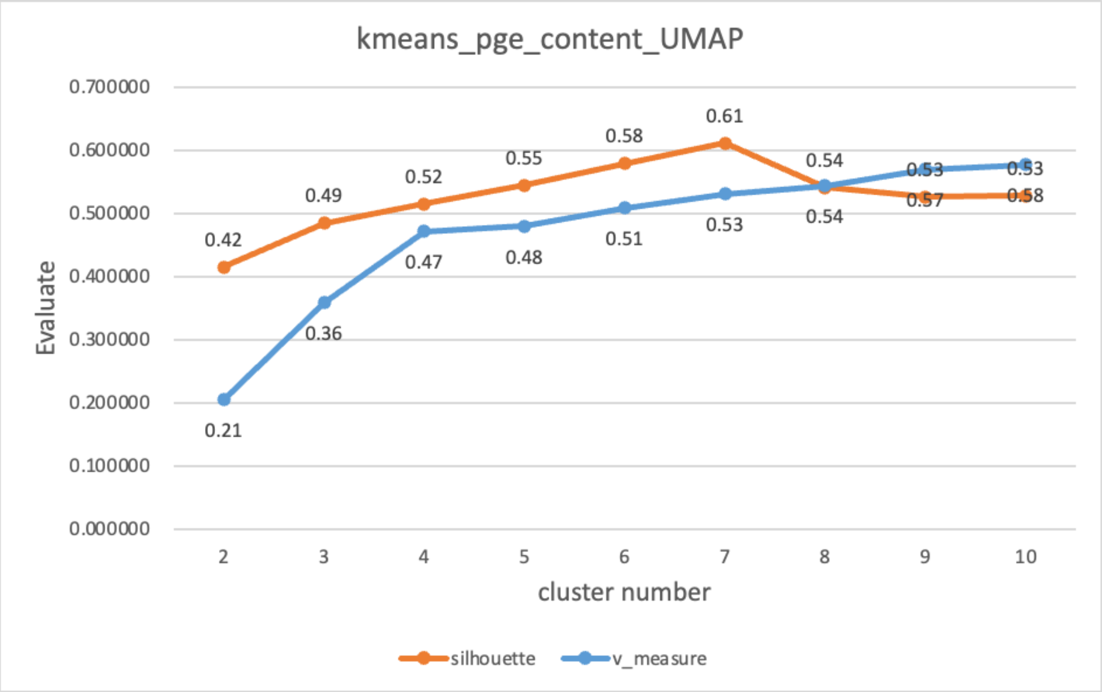
3.1 K-means

For two different datasets, we build different models by varying the number of k and calculate the V-measure and Silhouette Score to evaluate the models.

3.1.1 Dataset pge_content



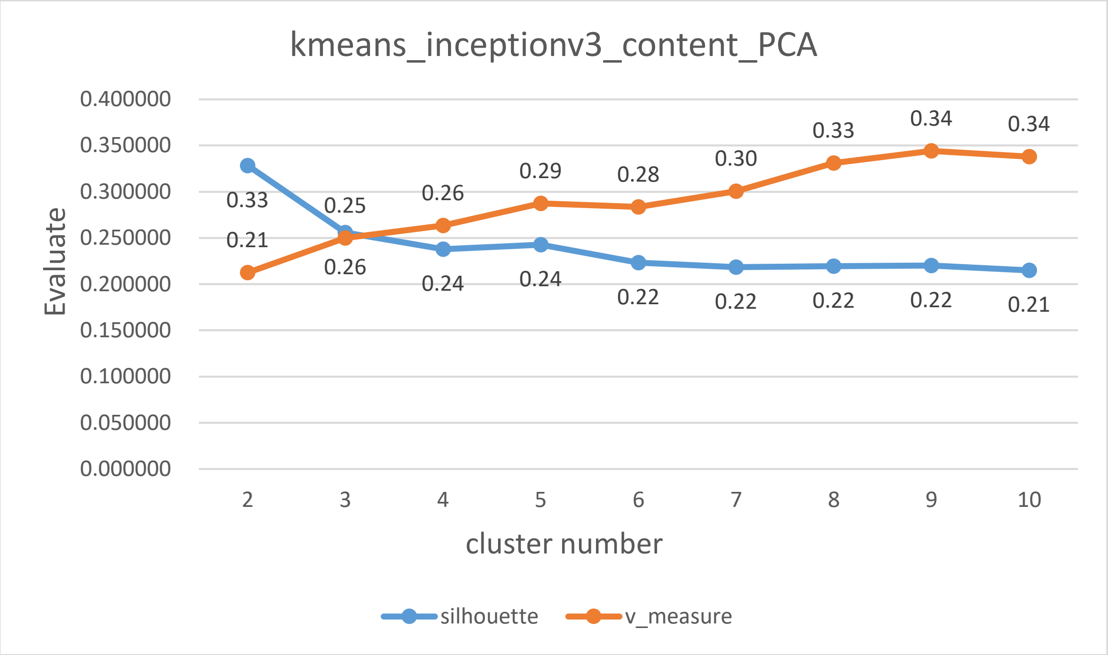
Graph3.1



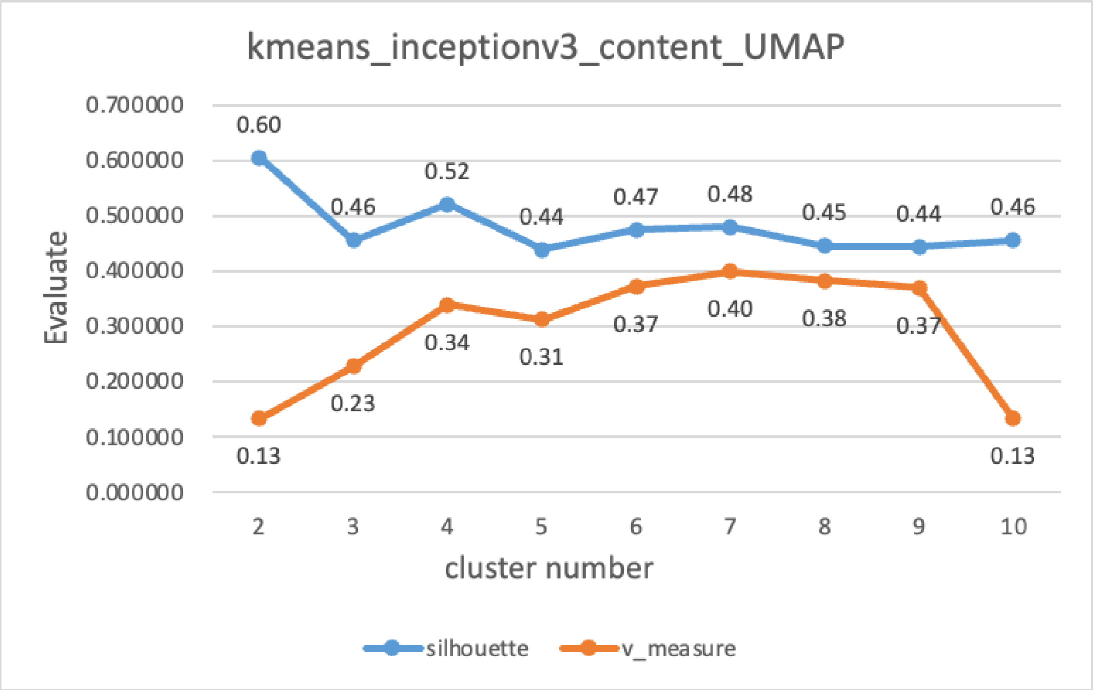
Graph3.2

The two graphs above present the changes in the plausibility of the model obtained by varying the number of k's for the data obtained by PCA and the data obtained by UMAP, respectively.

3.1.2 Dataset inceptionv3



Graph3.3

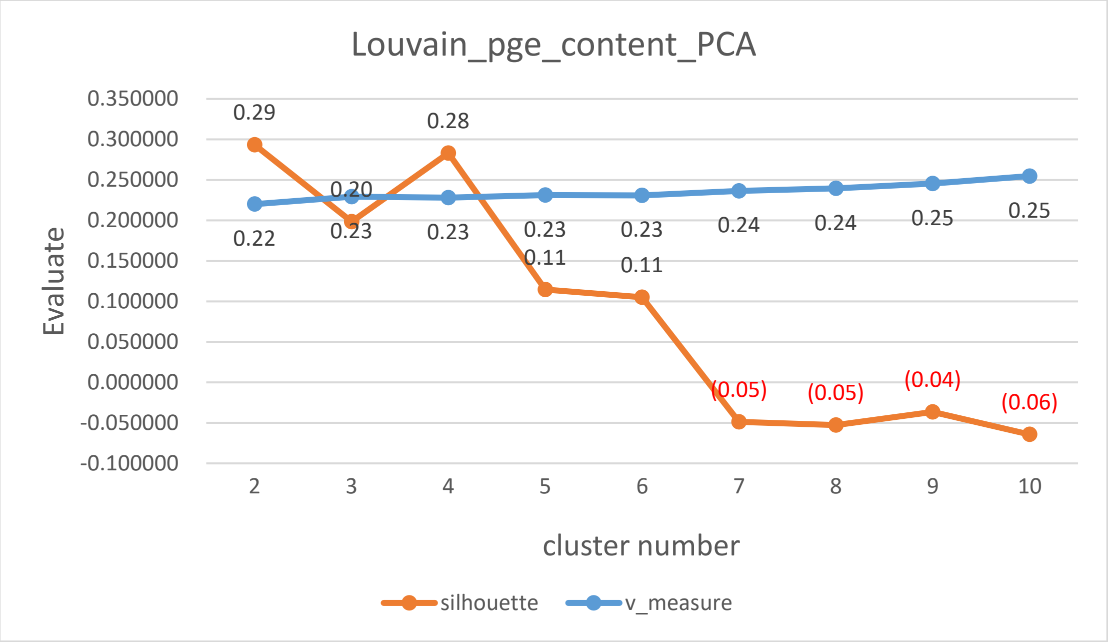


Graph3.4

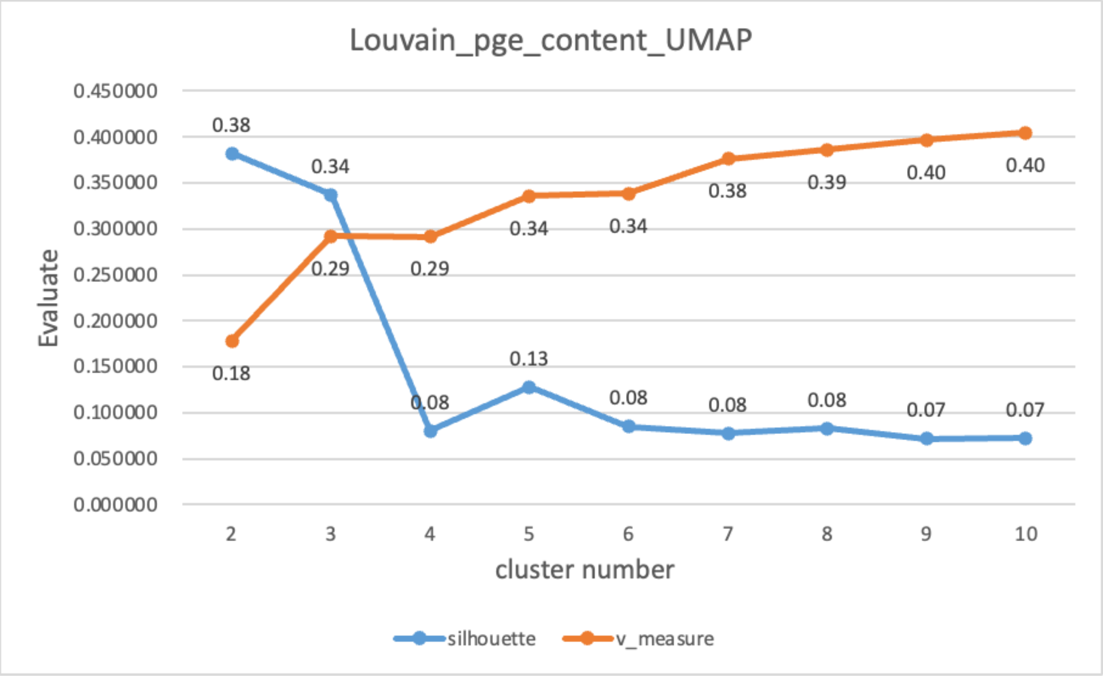
The two graphs above present the changes in the plausibility of the model obtained by varying the number of k's for the data obtained by PCA and the data obtained by UMAP, respectively.

3.2 Louvain Community Detection

3.2.1 Dataset pge_content



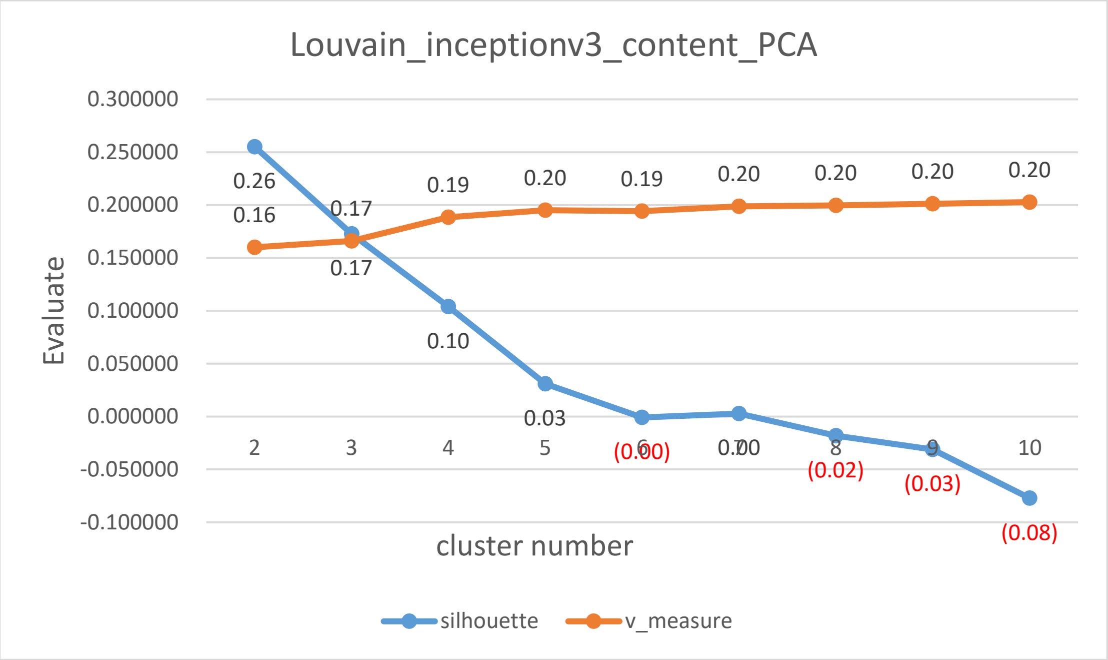
Graph3.5



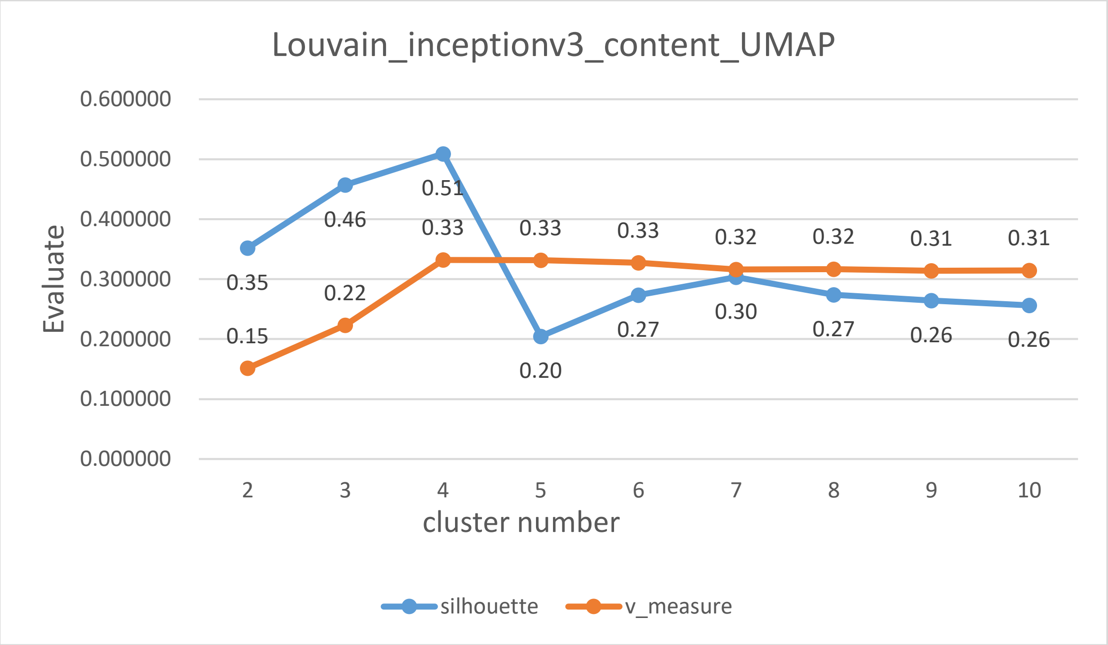
Graph3.6

The two graphs above present the changes in the reasonableness of the model obtained by varying the resolution of the data obtained by PCA and the data obtained by UMAP

3.2.2 Dataset inceptionv3



Graph3.7



Graph3.8

The two graphs above present the changes in the reasonableness of the model obtained by varying the resolution of the data obtained by PCA and the data obtained by UMAP

4.Results

K-means

PathologyGAN's optimal classification using K-means with PCA projection had a value of silhouette = 0.170619 (cluster number 3) and v-measure = 0.362236 (cluster number 9).

PathologyGAN's optimal classification using K-means with UMAP projection had a value of silhouette = 0.611286 (cluster number 7) and v-measure = 0.576979 (cluster number 10).

InceptionV3's optimal classification using K-means with PCA projection had a value of silhouette = 0.328390 (cluster number 2) and v-measure = 0.344154 (cluster number 9).

InceptionV3 optimal classification using K-means with UMAP projection had a value of silhouette = 0.604877 (number of clusters is 2) and v-measure = 0.398715 (number of clusters is 7).

Louvain Community Detection:

PathologyGAN's optimal classification using Louvain Community Detection with PCA projection had a value of silhouette = 0.293277 (cluster number 2) and v-measure = 0.254786 (cluster number 10).

PathologyGAN's optimal classification using Louvain Community Detection with UMAP projection had a value of silhouette = 0.381814 (2 clusters) and v-measure = 0.404838 (10 clusters).

InceptionV3's optimal classification using Louvain Community Detection with PCA projection had a value of silhouette = 0.255121 (2 clusters) and v-measure = 0.202834 (10 clusters).

InceptionV3 optimal classification using Louvain Community Detection with UMAP projection had a value of silhouette = 0.509153 (number of clusters is 4) and v-measure = 0.331952 (number of clusters is 4).
The following is a detailed analysis of the data.

PathologyGAN data sets

In the experiment, we randomly intercepted 2000 groups of data out of 5000 groups for testing. It can be seen from the data that using the UMAP projection method gives better results than using the PCA projection method in the two datasets tested, **PathologyGAN** and Inceptionv3.

Where **PathologyGAN** has the following optimal classification results when using the K-means algorithm for clustering.:

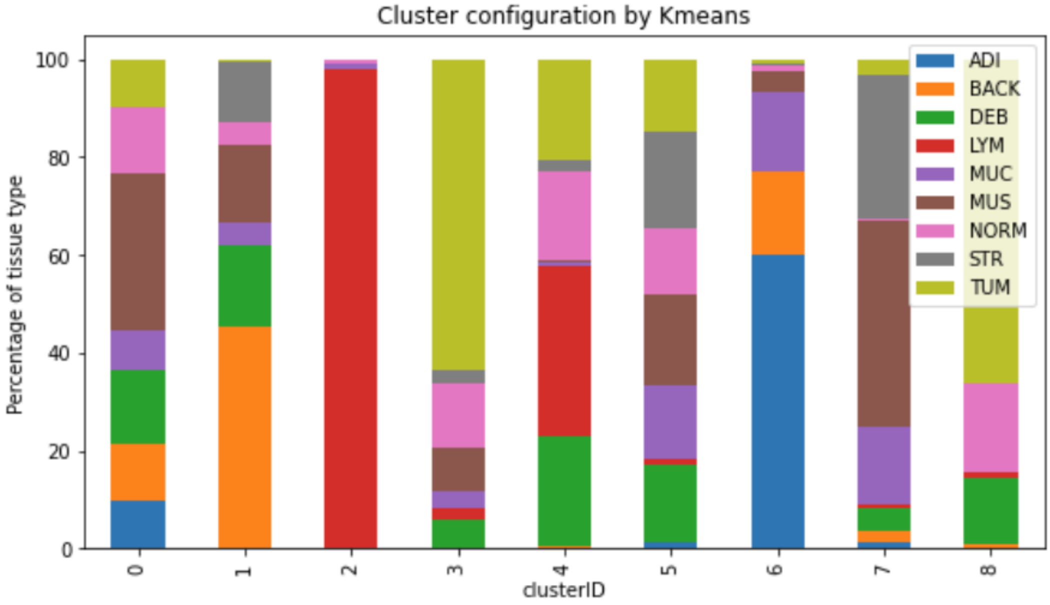


Figure 4.1 K-means PathologyGAN PCA best

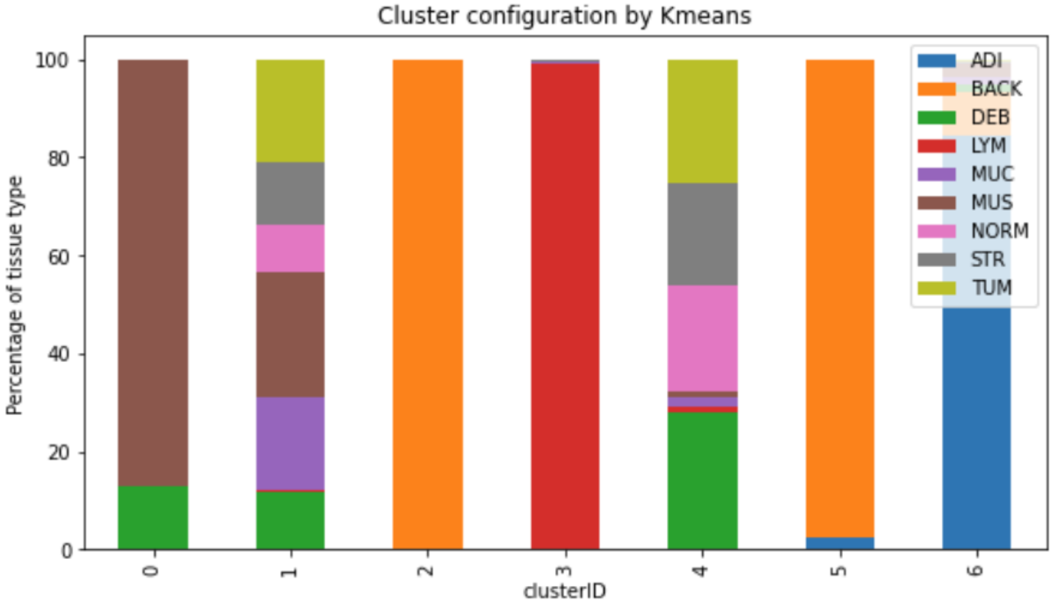


Figure 4.2 K-means PathologyGAN UMAP best

It can be concluded from the two images above that the clustering results in Figure 4.1 are not ideal. The TUM representing colorectal adenocarcinoma epithelium is present in each cluster. Figure 4.2 shows a much better result, with TUM only present in clusters 1 and 4.

Similar results were found when using the Louvain Community Detection method.

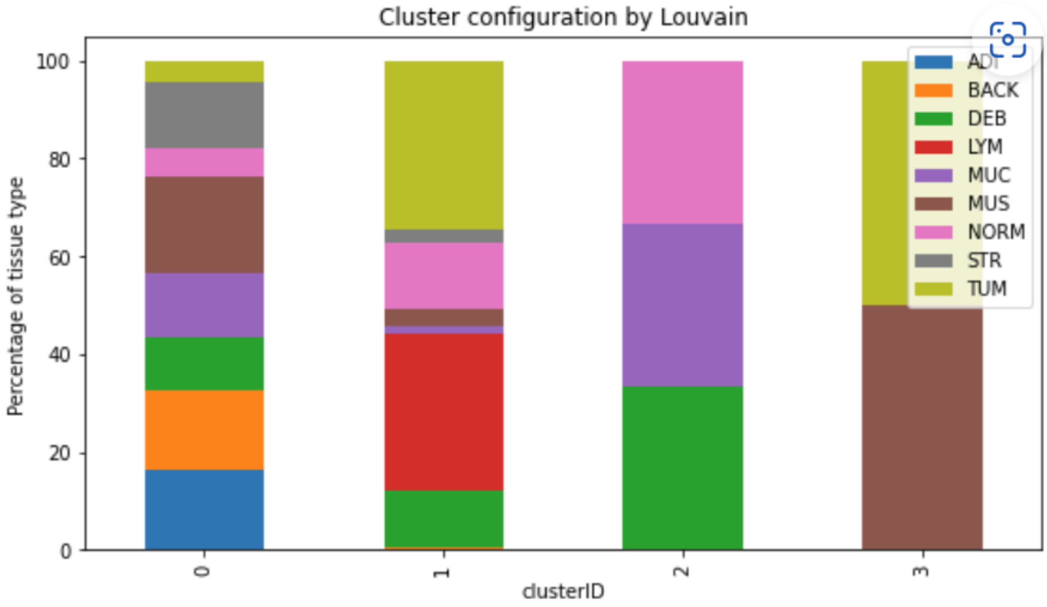


Figure 4.3 Louvain PathologyGAN PCA best

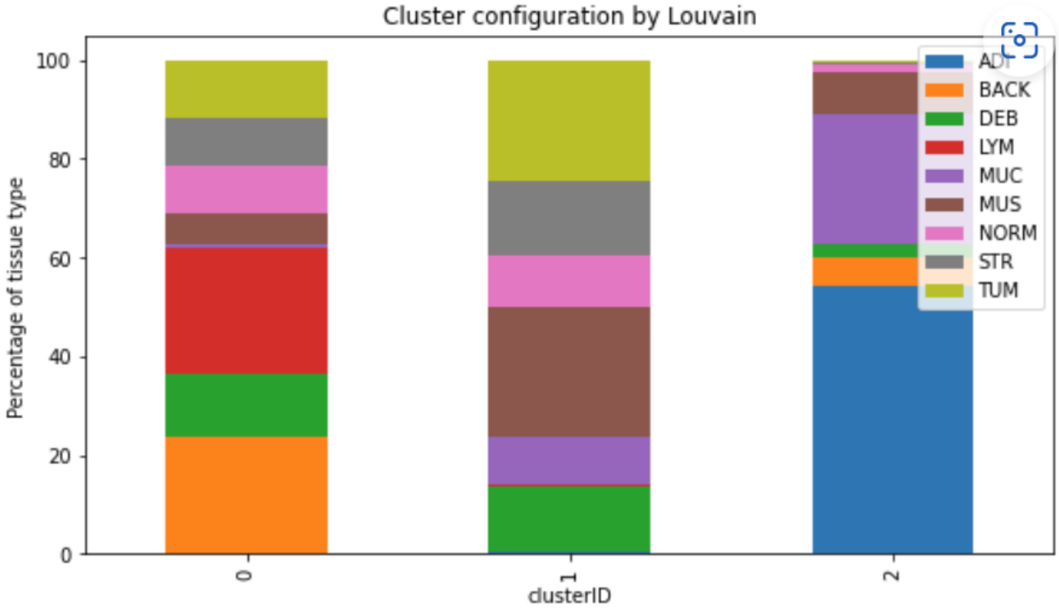


Figure 4.4 Louvain PathologyGAN UMAP best

Compared to the K-means algorithm, the classification results for Louvain Community Detection in this dataset are somewhat worse.

InceptionV3 data sets

When using PCA projection, the K-means algorithm achieves the best classification results when the cluster is 9.

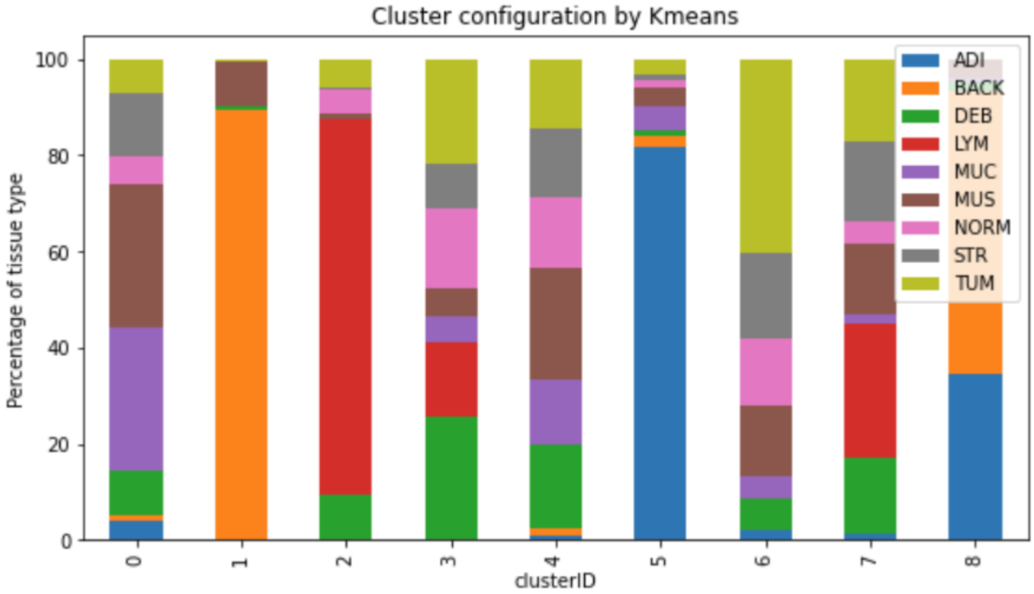


Figure 4.5 K-means InceptionV3 PCA best

And when using the UMAP projection, the K-means algorithm achieves the best classification results when the cluster is 7.

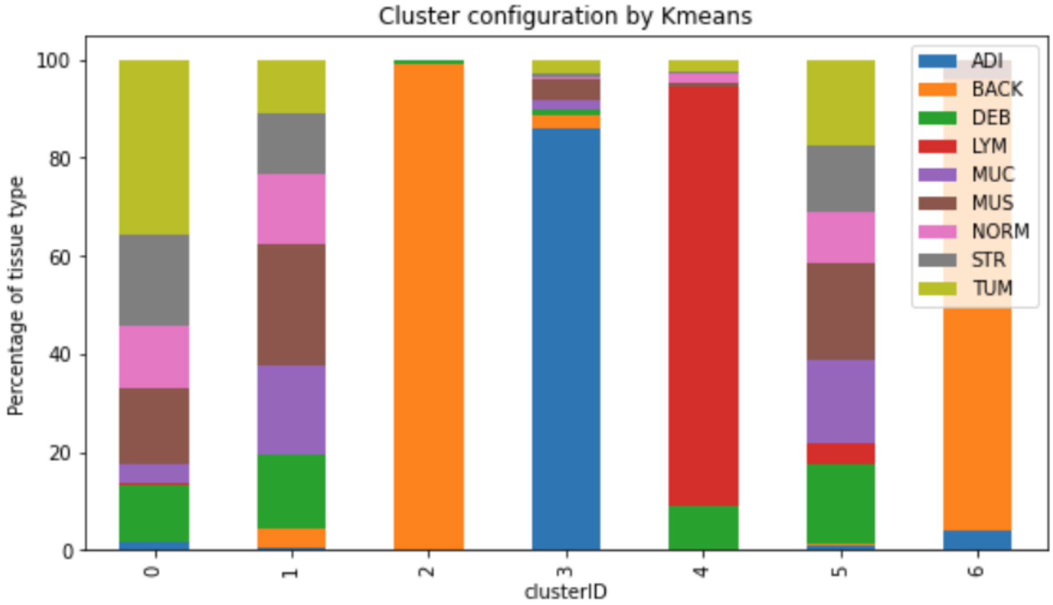


Figure 4.6 K-means InceptionV3 UMAP best

The Louvain Community Detection algorithm also performs slightly worse when using PCA projections.

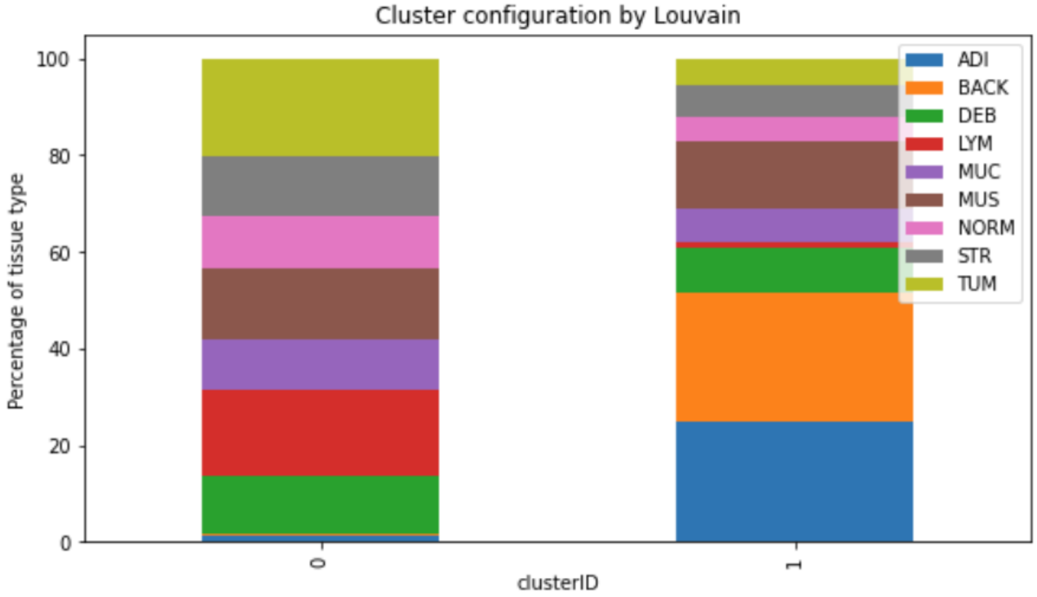


Figure 4.7 Louvain InceptionV3 PCA best

In contrast, the classification accuracy of the Louvain Community Detection algorithm is significantly improved when using the UMAP projection.

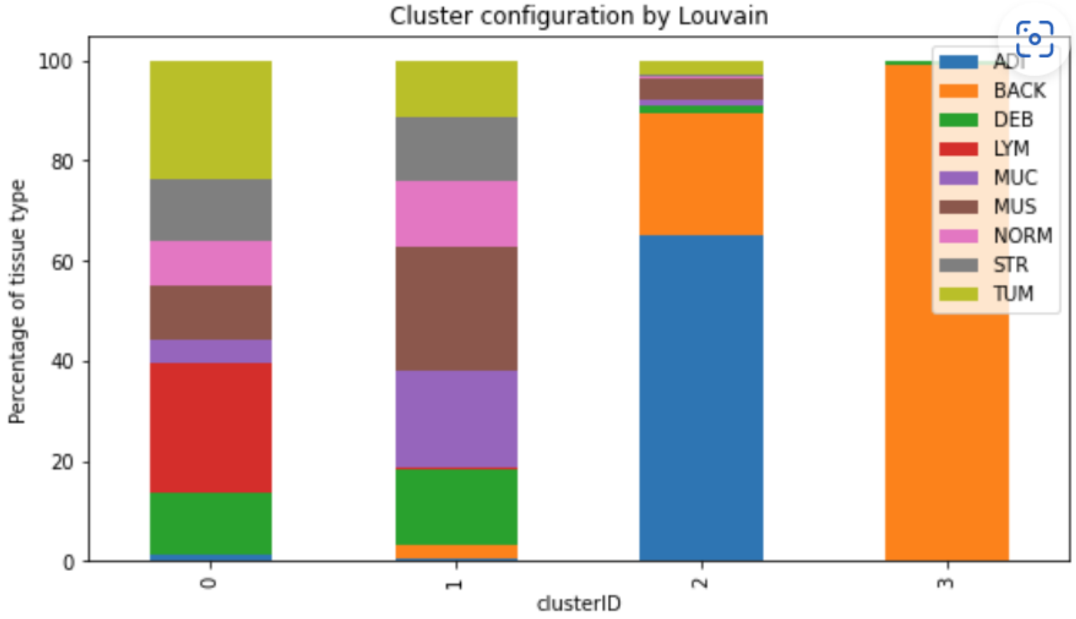


Figure 4.8 Louvain InceptionV3 UMAP best

5.Conclusion

As the aim of this experiment is tissue patch clustering, accuracy should be a priority. Therefore, the optimal model is chosen to minimise false positives. By analysing the results of the K-means and Louvain Community Detection algorithms, we found that the Louvain Community Detection algorithm appears to be more prone to false positives. And K-means ensures that the TUM is divided into a small number of clusters.

Analysis of the values of silhouette and v-measure also shows that both K-means and Louvain Community Detection have significantly higher values of silhouette and v-measure when using the UMAP projection than the PCA projection.

And the optimal number of clusters for Louvain Community Detection is concentrated at 2-4.When the number of clusters increases, Louvain Community Detection is unable to accurately separate TUM. on the other hand，The K-means algorithm, is significantly better in terms of stability of classification. When the number of clusters increases, K-means also ensures that TUM does not appear in half of the clusters, whereas the Louvain Community Detection algorithm has TUM in almost all of the clusters. In addition, we found that when using the Louvain Community Detection algorithm with PCA projection, the silhouette value decreases rapidly and even becomes negative when the number of clusters is too large. However, the v-measure values did not change much. By analysing the number of members in each cluster we found that several of the clusters split by Louvain Community Detection had very few members. The number of members is in the single digits to ten digits, which is a large difference from the total sample size of 2000.

From the above data we can conclude that in this experiment the classification effect using the UMAP projection is significantly better than the PCA projection. The K-means algorithm is also more suitable for this experiment than Louvain Community Detection. In summary, we believe that the optimal model for this experiment is the K-means algorithm using the UMAP projection.