Case Study 1

Model selection for Clustering

Group 104

Garima Ashok Devnani (2653434D)
Pooja Anil Kurup (2708790K)
Rajnish Kumar Verma (2694285V)
Tejas Kundu (2647799K)

Contents

Table of Figures

Table of Tables

Chapter 1 Introduction

The goal of this case study is to compare different clustering techniques using two alternative representations generated from a colorectal tissue patch by defining the cluster's quality using both intrinsic and extrinsic metrics.

The aim is to execute clustering algorithms on supplied representations to determine how 'excellent' each clustering result is. The different clustering methods used are Kmeans and Hierarchical. The data used for representation is ResNet 50and Pathology Gen.

1.1 Colorectal Cancer

Colorectal cancer develops in the colon or rectum. Depending on where they begin, these malignancies are sometimes known as colon cancer or rectal cancer. Since they share many characteristics, colon cancer and rectal cancer are sometimes lumped together.

In the case study, we are using 5000 colorectal cancer tissue patches. It is then segregated into 9 tissue types consisting of Adipose (ADI), Background (BACK), Debris (DEB), Lymphocytes (LYM), Mucus (MUC), Smooth muscle (MUS), Normal colon mucosa (NORM), Cancer-associated stroma (STR) and Colorectal adenocarcinoma epithelium (TUM). Reference Fig 1.

1.2 Data Projection

Every high-dimensional data point is projected onto an appropriate subspace(lower-dimensional space) in such a manner that the distances between the points are roughly preserved.

In the case study, we are using PCA and UMAP for Data Projection.

1.2.1 PCA (Principal Component Analysis)

PCA (Principal Component Analysis) is a well-known "unsupervised" dimensionality reduction approach. It works by determining the hyperplane that is nearest to the data and then projecting the data onto the hyperplane while keeping the majority of the data set's variance. (Sivarajah, 2020)

The Principal Components axis is the axis that explains the most variance in the training set. The second principal component is the axis perpendicular to this axis. As the dimensions increase, PCA will uncover a third component orthogonal to the previous two components, and so on; however, for visualization reasons, we always keep to two or three principal components. (Sivarajah, 2020)

1.2.2 UMAP (Uniform Manifold Approximation and Projection

UMAP is a nonlinear dimensionality reduction approach that is especially helpful for displaying clusters or groupings of data points concerning their relative proximities. It may be used directly on sparse matrices, removing the requirement for any prior pre-processing techniques such as PCA or Truncated SVD (Singular Value Decomposition). As a result of the increased processing speed, visualization is likely to be enhanced. (Sivarajah, 2020)

1.3 Data Set

A group of data items that a computer may treat as a single entity for analytic and predictive purposes. This means that the information obtained should be homogenous and understandable to a machine, which does not perceive information in the same way that humans do. It is vital to preprocess the data by cleaning and completing it, as well as annotate the data by adding relevant computer-readable tags once it has been collected. There are 5,000 images of different dimensional feature spaces.

1.3.1 ResNet50

ResNet50 is a 50-layer deep convolutional neural network. A pre-trained version of a network trained with over 1 million pictures from the ImageNet database may be loaded. Pre-trained networks can classify photos into 1000 different item categories, including humans, vehicles, pencils, and a variety of animals, etc. Therefore, the network has learned a comprehensive feature representation of a diverse set of pictures.

1.3.2 Pathology GAN (PGE)

Tumor histopathological pictures reveal a plethora of data about how the tumor develops and interacts with its surroundings. Greater knowledge of the tissue phenotype in these pictures might reveal new determinants of cancer's underlying pathological processes, improving diagnostic and treatment choices. While developments in deep learning are excellent for attaining these aims, their implementation is hampered by the expensive expense of labeling patient data with high-quality labels. Deep generative models with unsupervised learning, particularly those with representational learning features, offer an alternate method for better understanding the phenotype of malignant cells and capturing tissue morphology. (Quiros, et al., 2019)

Chapter 2 Methodology

2.1 Clustering Methods

Clustering is the task of categorizing a large set of data points into groups so that data points in the same group are more identical to other data points in the same grouping than data points in other groups. Segregate groups with similar features and assign them to a cluster identical or like the clusters. (Kaushik, 2016)

We apply K means and Hierarchical (Agglomerative) clustering in the case study along with ResNet 50 and Pathology GAN data.

2.1.1 Kmeans Clustering

Kmeans clustering is a basic and widely used unsupervised machine learning technique. The unmanned algorithm infers from the dataset using only the input vector and no reference to known or labeled outcomes. “Andrey Bu, who has more than 5 years of machine learning experience and currently teaches people his skills, says that “the objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset””. (Garbade, 2018)

The Kmeans algorithm identifies the number k of centroids and assigns each data point to the closest cluster while keeping the centroids as small as possible. Kmeans "means" refers to the average of the data which is finding the centroid. (Garbade, 2018)
Reference Fig3.

2.1.1.1 Elbow Method

The ideal number of clusters into which the data may be grouped is a critical stage for any unsupervised technique. One of the most prominent approaches for determining the ideal value of k is the Elbow Method.

2.1.2 Hierarchical (Agglomerative) Clustering

Hierarchical clustering, also known as hierarchical cluster analysis, is a method that organizes related items into clusters. Endpoints are cluster groupings; each cluster is distinct from the others, and the objects in each cluster are similar. Distance matrices or raw data can be used to accomplish hierarchical clustering. When you provide raw data, the program calculates the distance matrix automatically in the background. (Bock)

Each observation is treated as a separate cluster in hierarchical clustering. Two steps are performed in repetition. First, find the two clusters that are nearest to each other. Secondly, combine the two clusters that are the most similar. This iterative approach is repeated until all clusters are integrated.
Reference Fig 4.

2.1.2.1 Single Linkage

The smallest distance between a pair of observations in two clusters is known as single linkage (nearest neighbor). It can occasionally yield clusters in which observations from different clusters are closer together than data from the same cluster. These clusters might appear to be dispersed. (Sivarajah, 2020)

2.1.2.2 Ward Linkage

Ward's linkage is a hierarchical cluster analysis approach. The rise in the "error sum of squares" (ESS) after fusing two clusters into a single cluster is used to determine the linkage function, which specifies the distance between two clusters. Ward's Method aims to choose the successive

clustering phases so that the increase in ESS at each phase is kept to a minimum. (Meng, et al., 2018)

2.2 Performance Measurement

Assessing the quality of the clusters and strategies for measuring cluster performance The performance is measured using the Silhouette Score and the V-measure Score.

2.2.1 Silhouette Score

The silhouette score is used to assess the quality of clusters created using the clustering algorithm. Silhouette scores are calculated for each sample of different clusters. Determine the distance between each observation that belongs to all clusters to compute the silhouette score for each observation/data point. (Kumar, 2020)

$$s = (b - a) \div (max(a, b))$$

a=The average distance between the observation and all of the other data points in the same cluster. This distance is also known as the mean intra-cluster distance.

b=The average distance between the observation and the next closest cluster of data points. This is often referred to as the mean nearest-cluster distance.

The Silhouette Score ranges from **+1** to **-1.**

2.2.2 V-Measure Score

V-Measure was developed to evaluate any clustering technique's performance. (Gupta, 2019)

There are two components to determine the V-measure Score.

1. Homogeneity: A homogeneous clustering is one in which all data points in each cluster belong to the same class label. Homogeneity describes the clustering algorithm's proximity to perfection.

2. Completeness: A complete clustering is one in which all data points from the same class are grouped in a single cluster. The clustering algorithm's completeness defines how near it comes to this perfection.

The V-Measure Score ranges from **0** to +**1.**

Chapter 3 Result

3.1 Kmeans

We are deriving the optimal k based on Silhouette score. We obtained two outcomes for the feature set PGE for Silhouette Score: **0.307905** for PCA and **0.615687** for UMAP. Furthermore, we achieved **0.167148** for PCA and **0.607017** for UMAP in ResNet50. The optimal k value for PGE feature set using PCA is **2** and for UMAP is **7**. The optimal k value for ResNet 50 feature set using PCA is **4** and for UMAP it is **21**. The final members of Kmeans can be seen in Table3,4,5, and 6.

We obtained two PGE scores for the V measure: **0.455538** for PCA and **0.563564** for UMAP. In addition, we got two Vmeasure scores for ResNet 50: **0.547355** for PCA and **0.712187** for UMAP.

The final members of Kmeans clustering can be seen in Tables 3,4,5, and 6.

3.1.1 Elbow Method

We can see from the graphs below that calculating the precise k value from the curve using solely the Elbow technique is naïve and difficult to determine the exact value of the elbow. As a result, we calculate the silhouette score and v-measure to achieve the optimal k-value for both feature sets to gain a better measurement. Reference Fig 5, Fig 6, Fig 7, and Fig 8.

3.1.2 Silhouette Score

Assuming that the ground truth labels aren't present, we'll use the silhouette score to determine which 'k' best fits the data. Reference Fig 9,10,11 and 12.

In real-world unsupervised learning applications, Silhouette Score is a realistic measure for determining the optimal number of clusters 'k' for a given dataset, as genuine labels are never provided.

3.1.3 V Measure Score

We are also determining the optimal 'k' using V-Measure to identify the 'k' with the proper cluster quality, given that the ground truth labels are provided. Reference Fig 13,14,15 and 16.

3.1.4 Performance Score

Based on the optimal ‘k’ we get both the metric scores. All the respective metric scores for K-means clustering are provided below.

Metrics		PGE	ResNet 50
Silhouette Score	**PCA**	0.307905	0.167148
	UMAP	0.615687	0.607017
VMeasure	**PCA**	0.455538	0.547355
	UMAP	0.563564	0.712187

Table 1. KMeans Metrics Score

3.2 Hierarchical

Due to the structure of the feature set, we are modifying the linkage = 'ward' and joining the nearest branch depending on the variance of each cluster.

We obtained two outcomes for the feature set PGE for Silhouette Score: **0.31998** for PCA and **0.615047** for UMAP. Furthermore, we achieved **0.151562** for PCA and **0.610128** for UMAP in ResNet50. The optimal k value for PGE feature set using PCA is **2** and for UMAP is **7**. The optimal k value for ResNet 50 feature set using PCA is **4** and for UMAP it is **5**.

We obtained two PGE scores for the V measure: **0.450976** for PCA and **0.566360** for UMAP. In addition, we got two Vmeasure scores for ResNet 50: **0.576000** for PCA and **0.729171** for UMAP. The optimal k value for PGE feature con set using PCA is 26 and for UMAP is 18. The optimal k value for ResNet 50 feature set using PCA is 12 and for UMAP it is 15.

The final members of Hierarchical Clustering can be seen in Tables 7,8,9, and 10.

3.2.2 Silhouette Score

Assuming that the ground truth labels aren't present, we'll use the silhouette score to determine which 'k' best fits the data. Reference Fig. 17,18,19 and 20.

3.2.3 V Measure Score

We are also determining the optimal 'k' using V-Measure to identify the 'k' with the proper cluster quality, given that the ground truth labels are provided. Reference Fig. 21,22,23 and 24.

3.2.4 Performance Score Linkage = ‘Ward’

All the metric scores for Hierarchical clustering are included below.

Metrics		**PGE**	**ResNet 50**
Silhouette Score	**PCA**	0.31998	0.151562
	UMAP	0.615047	0.610128
VMeasure	**PCA**	0.450976	0.576000
	UMAP	0.566360	0.729171

Table 2. Hierarchical Clustering Linkage = ‘Ward’ Metrics Score

Overall, we can see that when we cluster with respect to the variance of each cluster, i.e., when linkage = 'ward,' we receive higher performance ratings and more acceptable clusters.

3.3 Graphs

Each feature set has two projections (UMAP and PCA), and each projection has two graphs (Kmeans and Hierarchical). They visualize the tissue type percentage for the different clustering techniques. Reference Fig 25,26,27 and 28.

Chapter 4 Discussion

We utilized the PGE and Resnet50 feature sets, as well as their respective PCA and UMAP projections, to cluster data using two methods. The two clustering methods used are K -means Clustering and Hierarchical Clustering (Agglomerative). To select the best cluster 'k,' we attempted the elbow approach. However, since the elbow point is a simplistic way of determining the best 'k' value. The silhouette scores were used to determine the best 'k' for both feature sets.

We also noted that the results for the PGE dataset were consistent across both projections. However, we noticed a slight difference in the UMAP projections between the two clustering techniques for the ResNet 50 feature set.

When the ground truth labels are present, the values of cluster 'k' obtained using the v-measure performance metric are superior to the values acquired using the silhouette score performance metric. However, because ground truth labels are seldom accessible in real-world situations, it is preferable to use the silhouette score for evaluating the success of unsupervised machine learning approaches such as clustering.

References

1. Anon., 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS.*
2. Fiori, L., 2020. *K-Means Clustering using Python.* [Online]
   Available at: https://medium.com/@luigi.fiori.lf0303/k-means-clustering-using-python-db57415d26e6
3. Garbade, D. M. J., 2018. *Understanding K-means Clustering in Machine Learning.* [Online]
   Available at: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
4. Gupta, A., 2019. *ML | V-Measure for Evaluating Clustering Performance.* [Online]
   Available at: https://www.geeksforgeeks.org/ml-v-measure-for-evaluating-clustering-performance/
5. Kaushik, S., 2016. *An Introduction to Clustering and different methods of clustering.* [Online]
   Available at: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/#h2_10
6. Kumar, A., 2020. *KMeans Silhouette Score Explained With Python Example.* [Online]
   Available at: https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam#
7. Meng, A., Li, K. & Hu, Y., 2018. *Agglomertive Hierarchical Clustering using Ward Linkage.* [Online]
   Available at: https://jbhender.github.io/Stats506/F18/GP/Group10.html
8. Quiros, A. C., Murray-Smith, R. & Yuan, K., 2019. PathologyGAN: Learning deep representations of cancer tissue. *Journal of Machine Learning for Biomedical Imaging.*
9. Sivarajah, S., 2020. *Dimensionality Reduction for Data Visualization: PCA vs TSNE vs UMAP vs LDA.* [Online]
   Available at: https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29
10. Anon., n.d. *ResNet 50.* [Online]
    Available at:https://www.mathworks.com/help/deeplearning/ref/resnet50.html#:~:text=ResNet%2D50%20is%20a%20convolutional,%2C%20pencil%2C%20and%20many%20animals
11. Bock, T. . What is Hierarchical Clustering?. . DISPLAYR.[Online]. [30 November 2021]. Available from: https://www.displayr.com/what-is-hierarchical-clustering/#:~:text=Hierarchical%20clustering%2C%20also%20known%20as,broadly%20similar%20to%20each%20other

Appendix 1 **Colorectal cancer tissue**

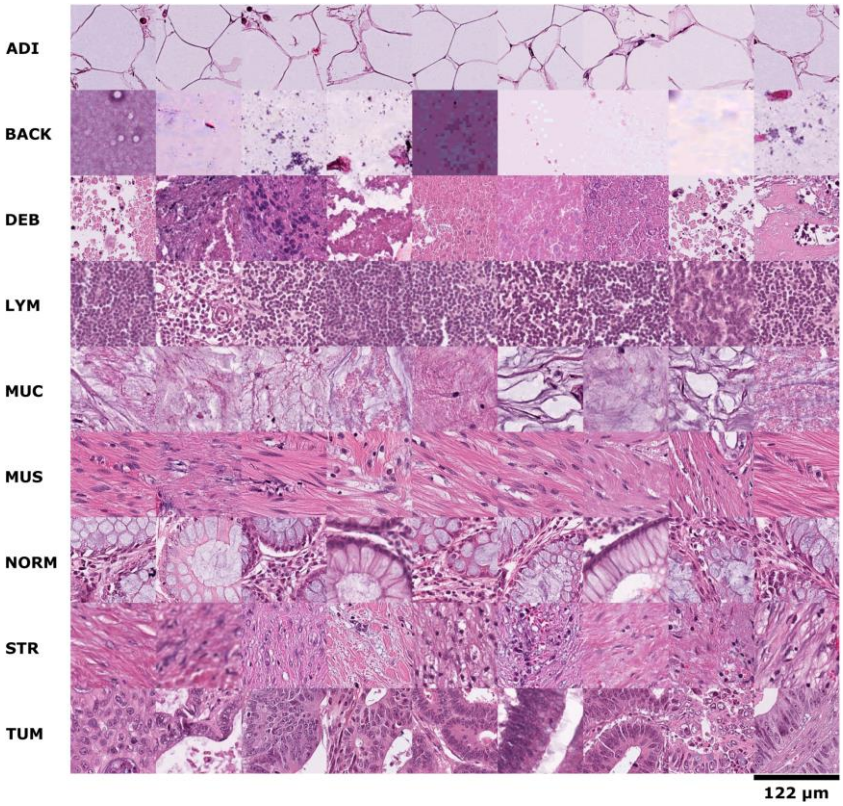


Fig 1 , Colorectal cancer tissue patches

Appendix 2 **PCA**

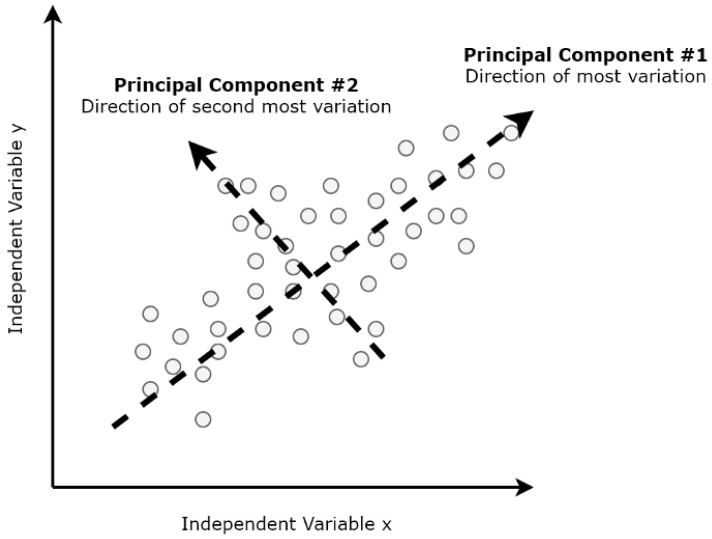


Fig 2, Principal Component Analysis

Appendix 3 **Kmeans Clustering**

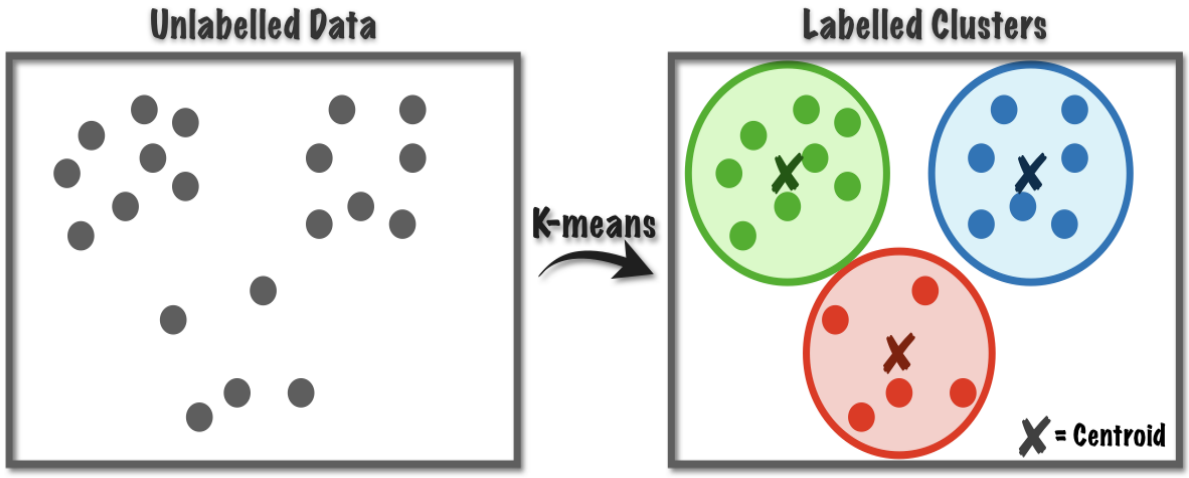


Fig 3, Kmeans

Appendix 4 **Hierarchical clustering**

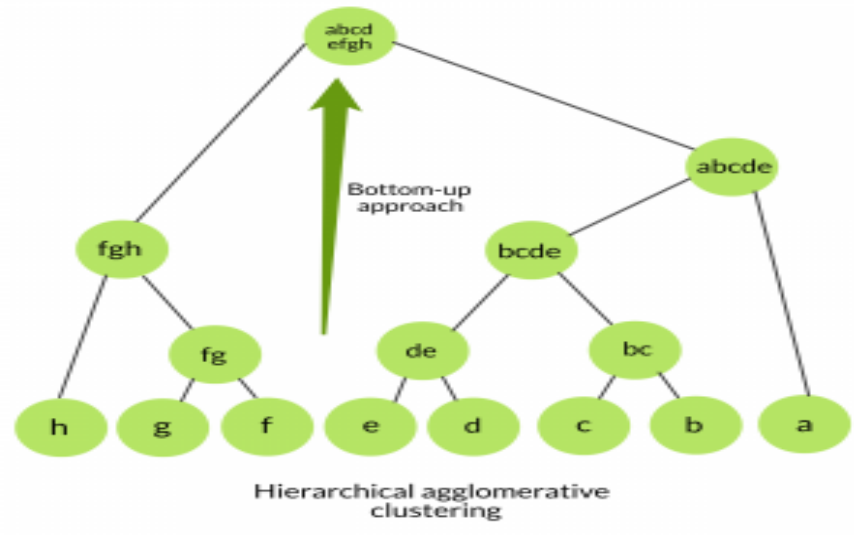


Fig 4, Hierarchical clustering

Appendix 5 **Elbow Method**

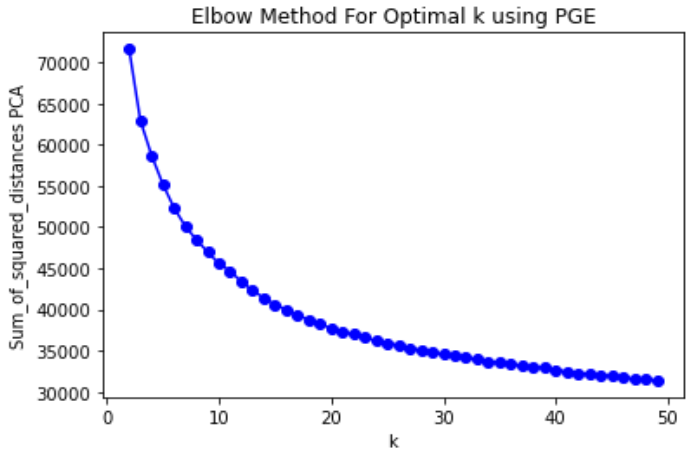


Fig 5, Elbow method for optimal k using PGE (PCA)

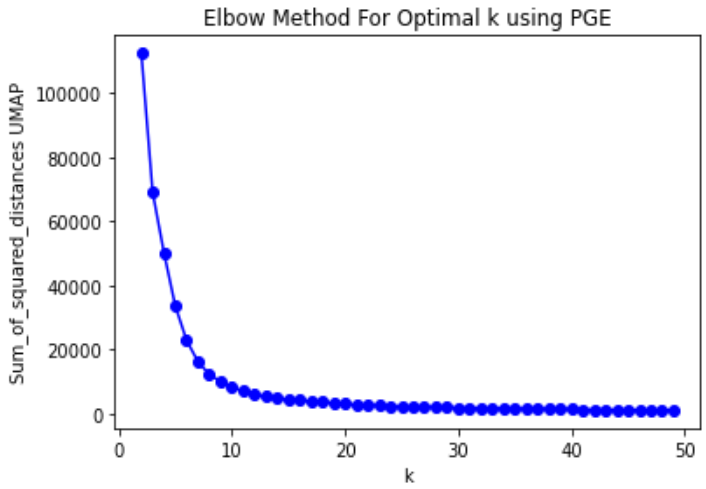


Fig 6, Elbow method for optimal k using PGE (UMAP)

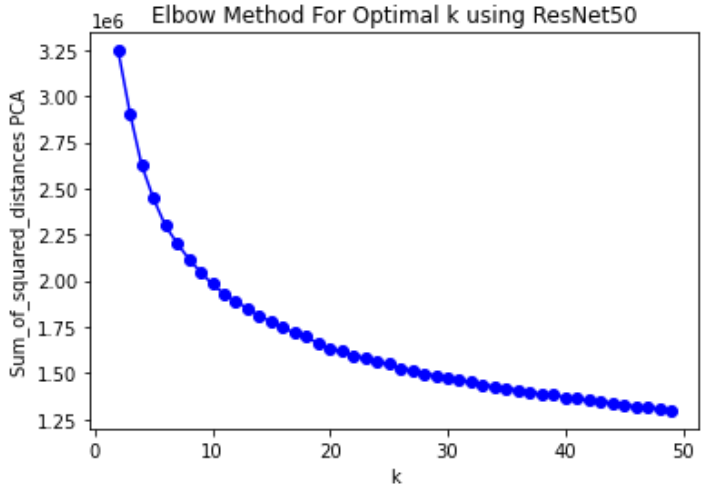


Fig 7, Elbow method for Optimal k using ResNet 50 (PCA)

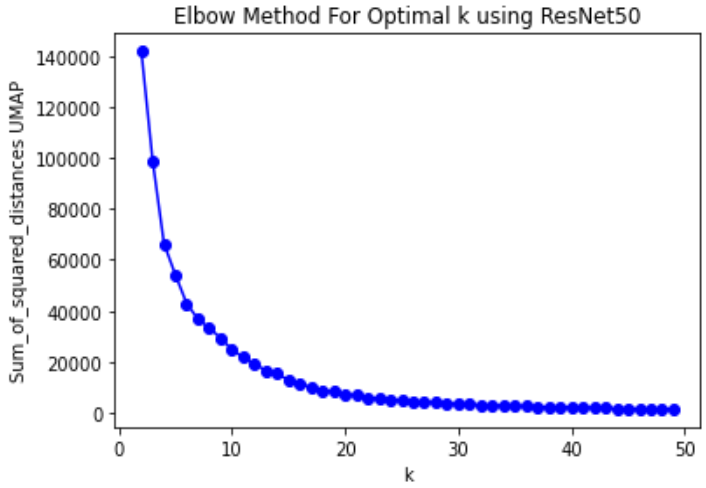


Fig 8, Elbow method for Optimal k using ResNet 50 (UMAP)

Appendix 6: **Kmeans Silhouette Score**

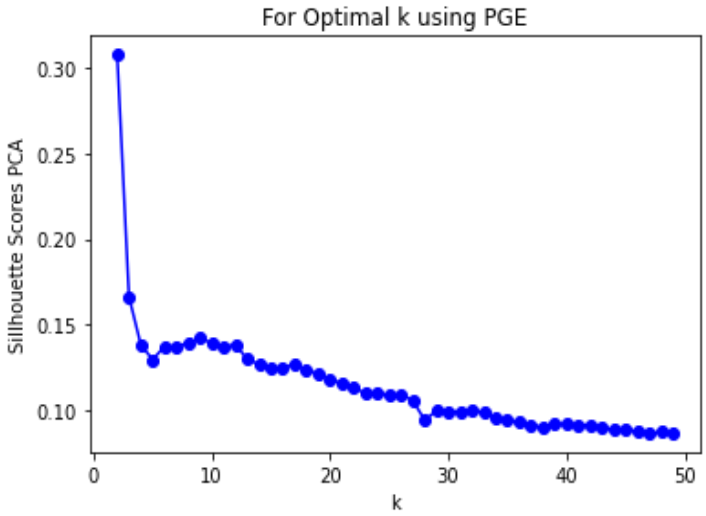


Fig 9, Silhouette Score (PCA) for Optimal k using PGE

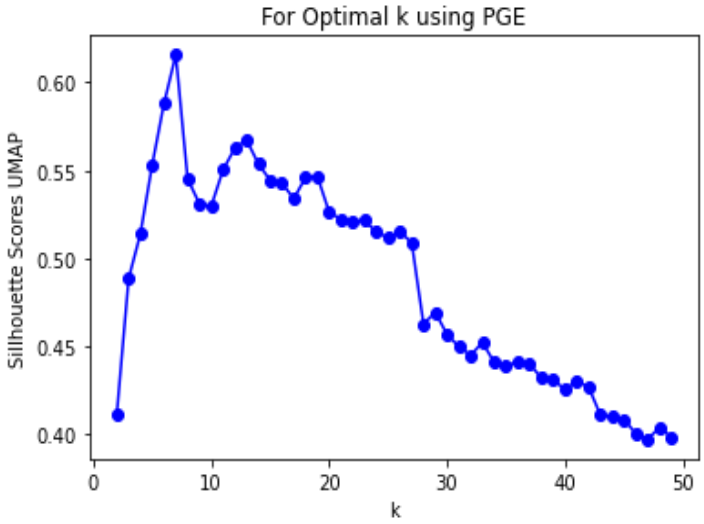


Fig 10, Silhouette Score (UMAP) for Optimal k using PGE

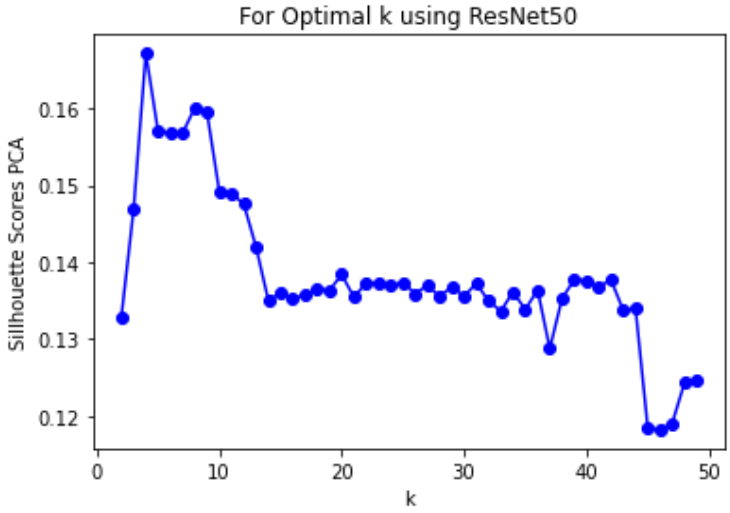


Fig 11, Silhouette Score (PCA) for Optimal k using ResNet 50

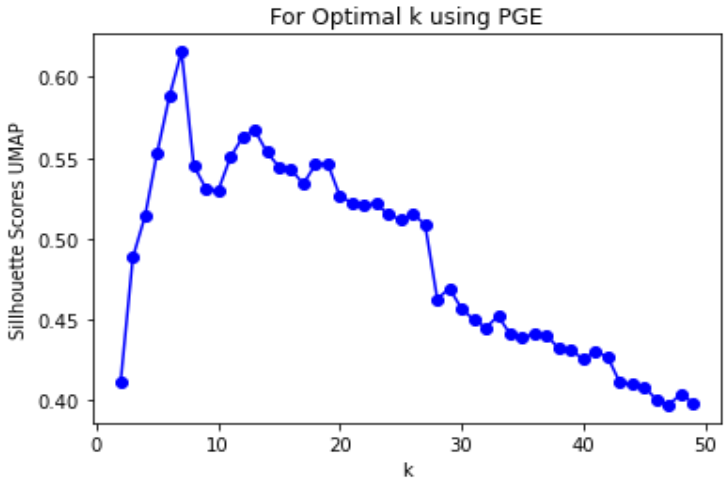
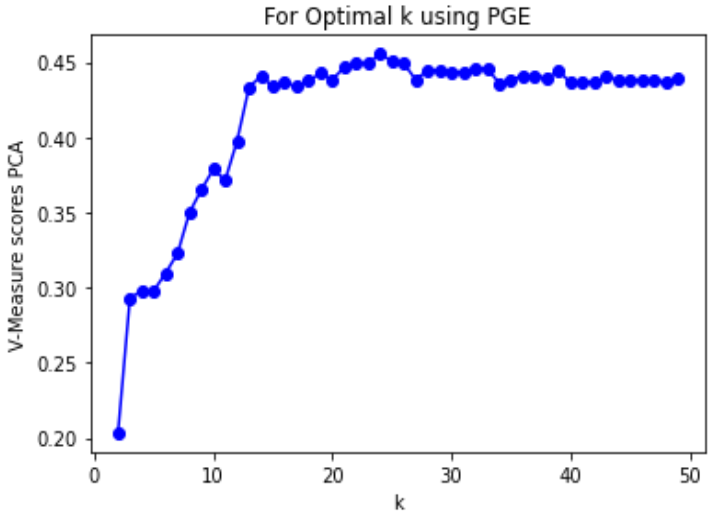


Fig 12, Silhouette Score (UMAP) for Optimal k using ResNet 50

Appendix 7 **Kmeans V-Measure**



to

Fig 13, V Measure score (PCA) for Optimal k using PGE

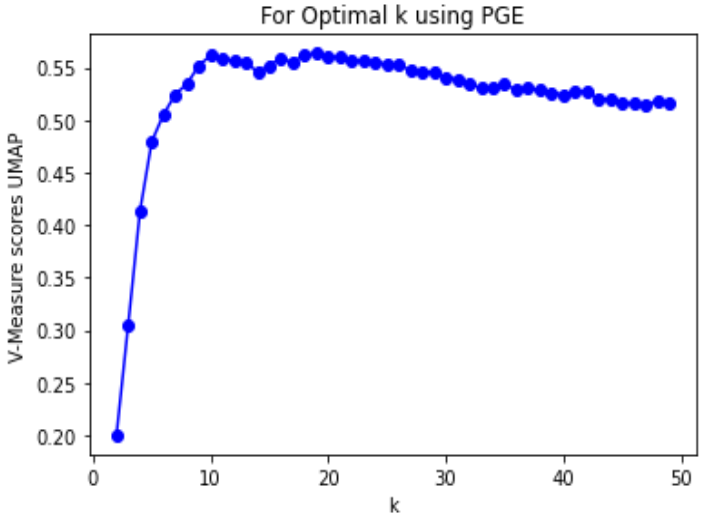


Fig 14, V Measure score (UMAP) for Optimal k using PGE

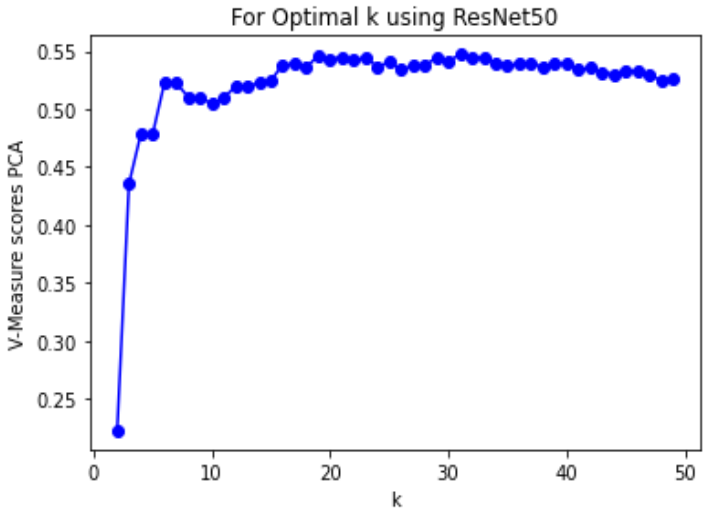


Fig 15, V Measure score (PCA) for Optimal k using ResNet 50

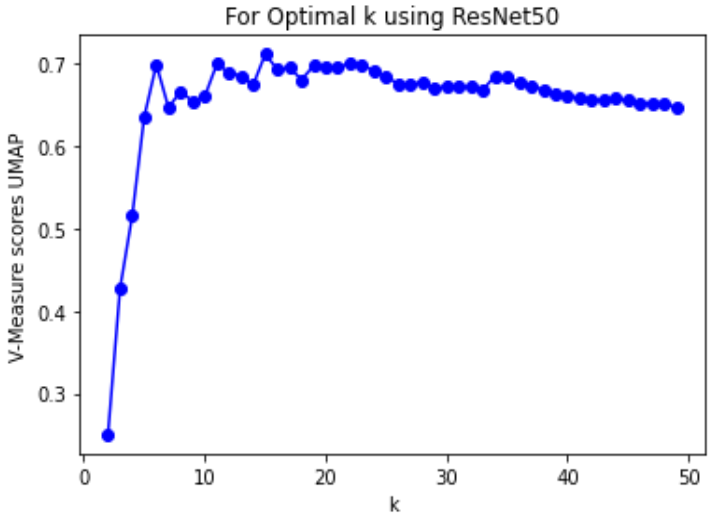


Fig 16, V Measure score (UMAP) for Optimal k using ResNet 50

Appendix 8 **Hierarchical Silhouette Score**

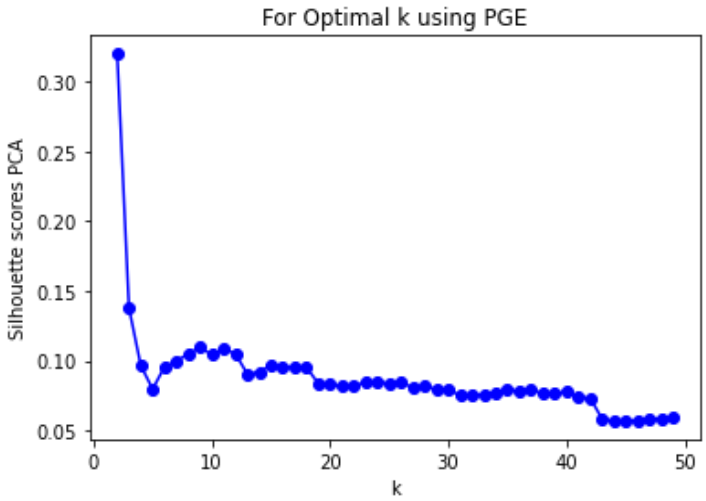


Fig 17, Silhouette Score(PCA) for Optimal k using PGE(H)

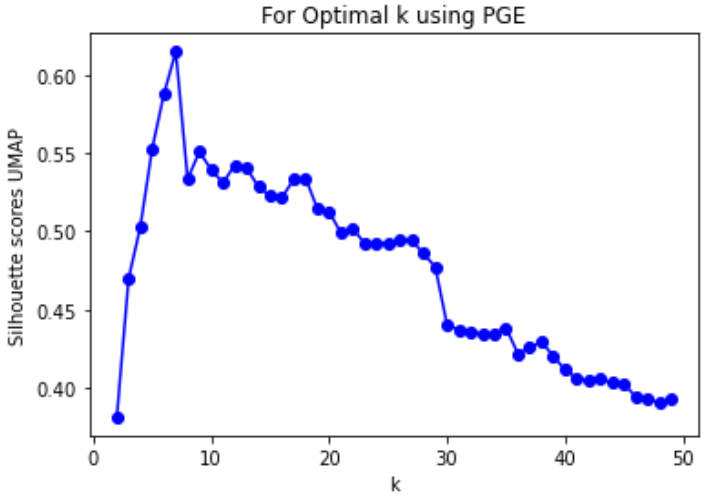


Fig 18, Silhouette Score(UMAP) for Optimal k using PGE(H)

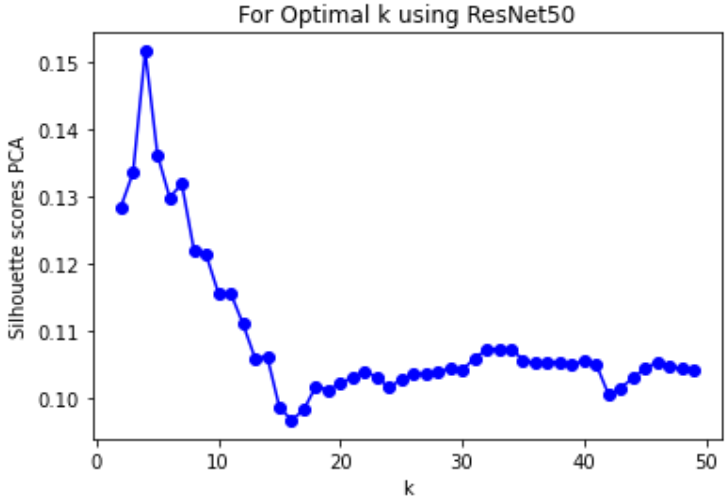


Fig 19, Silhouette Score(PCA) for Optimal k using ResNet 50(H)

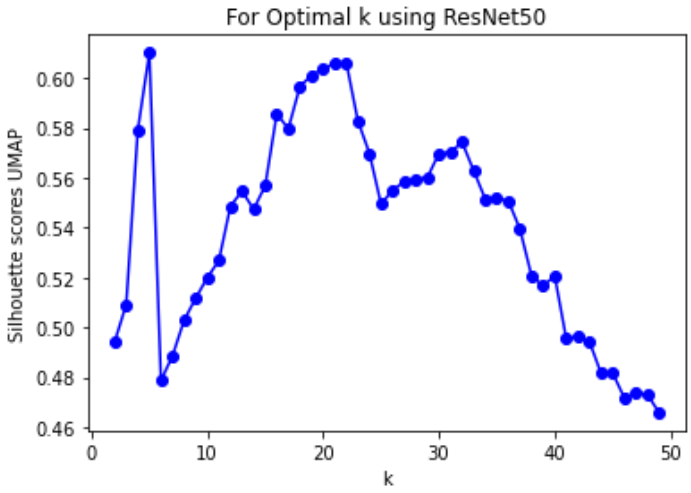


Fig 20, Silhouette Score(UMAP) for Optimal k using ResNet

Appendix 9 **Hierarchical Vmeasure Score**

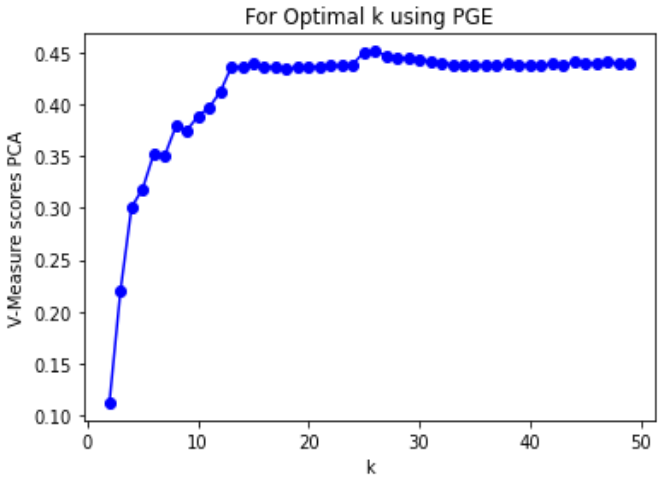


Fig 21, V Measure Score(PCA) for Optimal k using PGE(H)

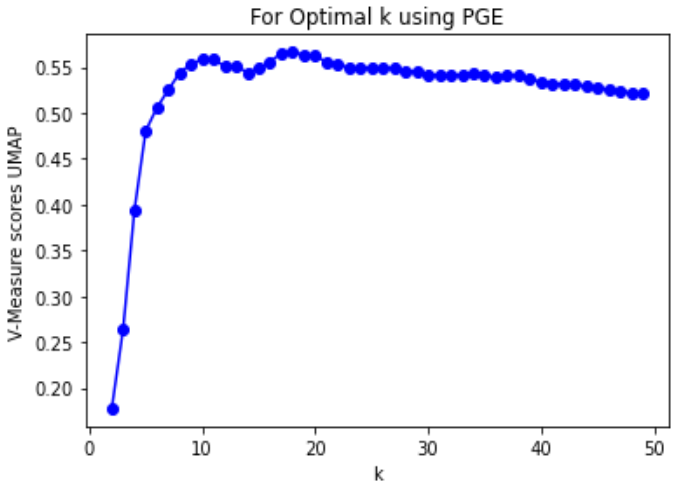


Fig 22, V Measure Score(UMAP) for Optimal k using PGE(H)

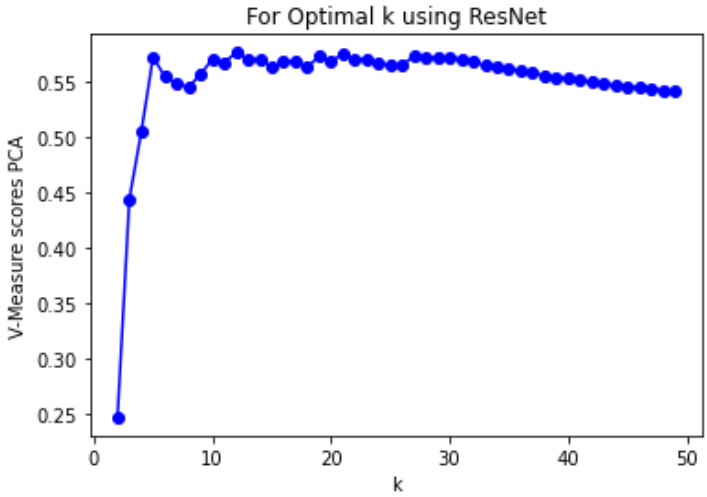


Fig 23, V Measure Score(PCA) for Optimal k using ResNet 50(H)

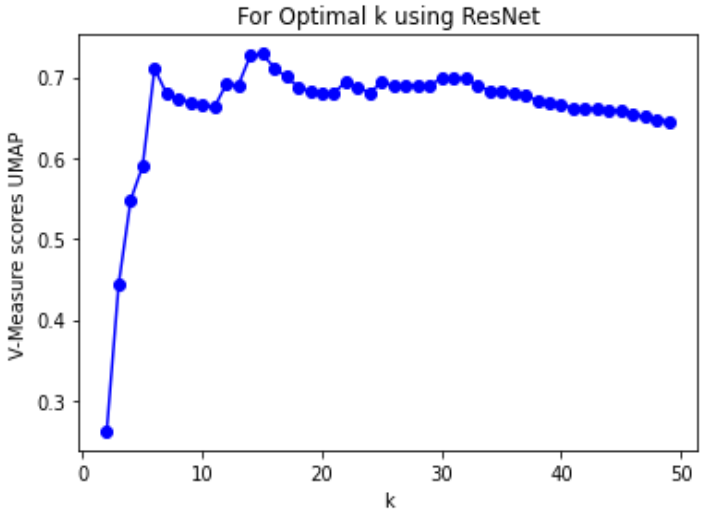


Fig 24, V Measure Score(UMAP) for Optimal k using ResNet 50(H)

Appendix 10 **Graphs**

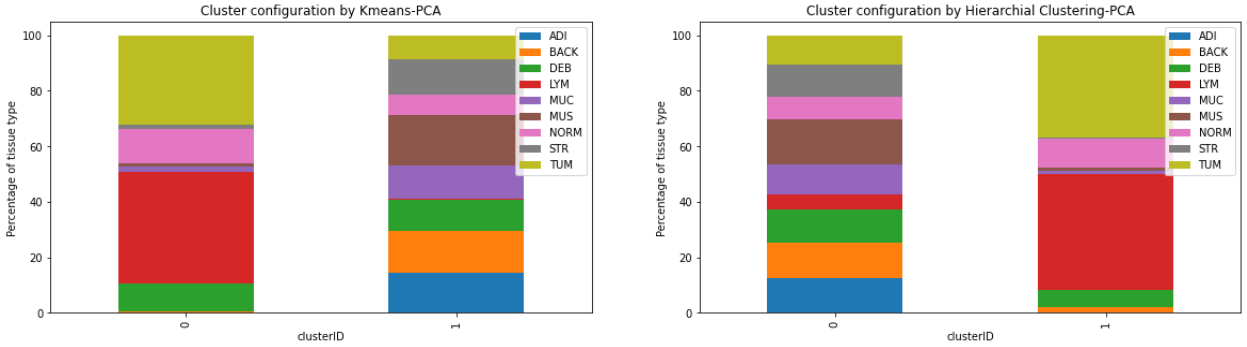


Fig 25. Cluster Configuration of PCA for PGE

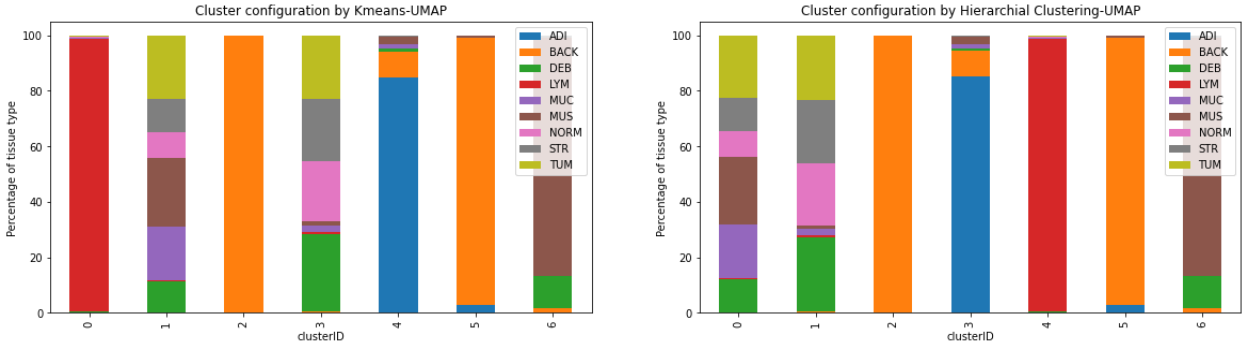


Fig 26. Cluster Configuration of UMAP for PGE

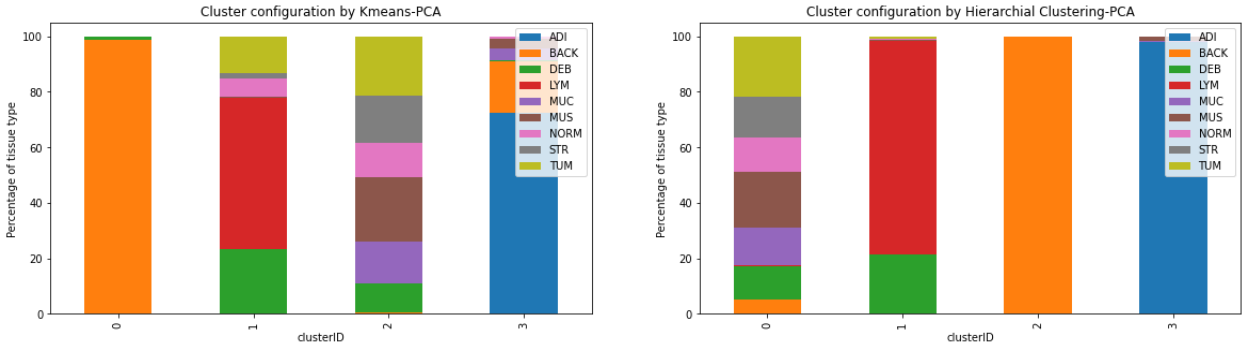


Fig 27. Cluster Configuration of PCA for ResNet50

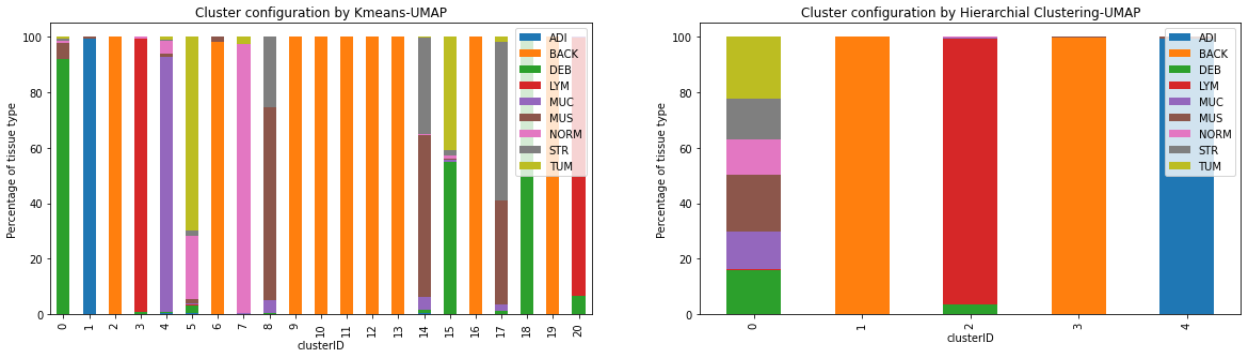


Fig 28. Cluster Configuration of UMAP for ResNet 50

Appendix 11 **Table**

Cluster Index	No of Members
0	1329
1	3671

Table 3. Kmeans PGE, PCA Assignment counts

Cluster Index	No of Members
0	415
1	1031
2	2825
3	729

Table 4. Kmeans ResNet 50, PCA Assignment counts

Cluster Index	No of Members
0	383
1	810
2	890
3	570
4	1047
5	518
6	782

Table 5. Kmeans PGE, UMAP Assignment counts

Cluster Index	No of Members
0	329
1	270
2	275
3	148
4	164

5	207
6	222
7	237
8	293
9	201
10	217
11	308
12	210
13	284
14	272
15	178
16	296
17	220
18	269
19	195
	205

Table 6. Kmeans ResNet 50, UMAP Assignment counts

Cluster Index	No of Members
0	463
1	293
2	511
3	236
4	30
5	121
6	184
7	197
8	233
9	164
10	66
11	209
12	87
13	204
14	239
15	105
16	453
17	76
18	201
19	173
20	242
12	104
22	104
23	183
24	40

25	82

Table 7. Agglomerative Clustering-PGE, PCA Assignment counts

Cluster Index	No of Members
0	381
1	534
2	761
3	340
4	401
5	359
6	434
7	314
8	617
9	321
10	281
11	257

Table 8. Agglomerative Clustering-ResNet 50, PCA Assignment counts

Cluster Index	No of Members
0	386
1	458
2	536
3	225
4	484
5	311
6	121
7	134
8	315
9	369
10	468
11	145
12	104
13	213
14	294
15	338
16	97
17	2

Table 9. Agglomerative Clustering PGE, UMAP Assignment counts

Cluster Index	No of Members
0	1186

1	387
2	581
3	210
4	31
5	91
6	566
7	105
8	34
9	60
10	530
11	744
12	20
13	442
14	13

Table 10. Agglomerative Clustering-ResNet 50, UMAP Assignment counts