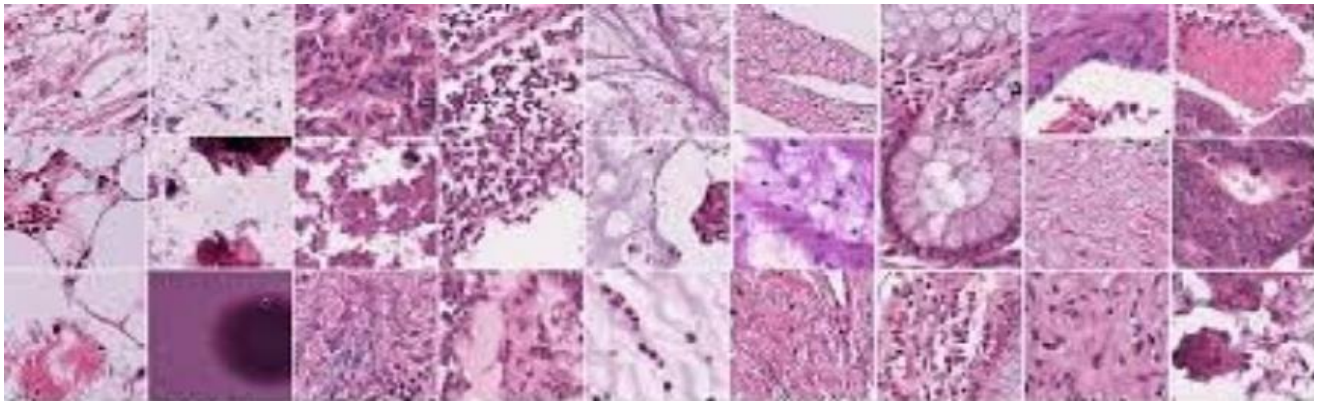


Case study 1: Model selection for Tissue Patch clustering

Sanjana Jammigumpala, Sarika Kamble, Yashika Rastogi & Laharika Tutica

2761536J@student.gla.ac.uk

2708819K@student.gla.ac.uk

2802810R@student.gla.ac.uk

2797248T@student.gla.ac.uk

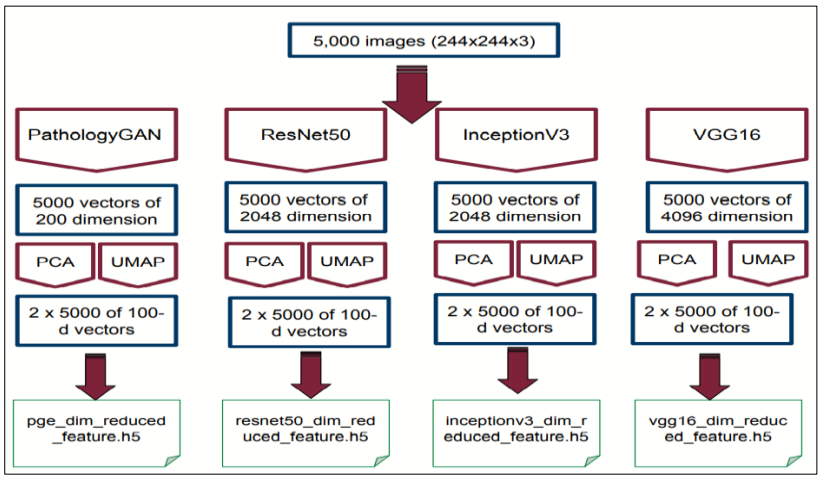
1. Introduction

1.1 Problem Statement

- Clustering is an unsupervised learning algorithm.
- Clustering is the task of diving data points into number of groups such that data points in same group are similar to other data points in the same group and dissimilar to data points in other groups.
- We have cancer dataset containing colorectal cancer tissue patches.
- Our task is to select clustering algorithm for cancer tissue types and assess the model performance.

1.2 Data

- We have 4 representations of the data as mentioned below.
  PathologyGAN
  ResNet50
  InceptionV3
  VGG16
- Below 2 dimensionality reduction methods are applied on these datasets to create array of size 5000 * 100.
  PCA
  UMAP



Tissue Types in Dataset :
There are 9 tissue types in the Dataset.

1. Adipose (ADI)
2. Background (BACK)
3. Debris (DEB)
4. Lymphocytes (LYM)
5. Mucus (MUC)
6. Smooth muscle (MUS)
7. Normal colon mucosa (NORM)
8. Cancer-associated stroma (STR)
9. Colorectal adenocarcinoma epithelium (TUM)

---

2. Methodology ,Parameter Searching and Evaluation of Clustering Algorithms

We will apply below 2 clustering algorithms to 2 representations - PythologyGAN and VGG16.

K-Means clustering algorithm

Hierachical - Agglomerative Clustering Algorithm

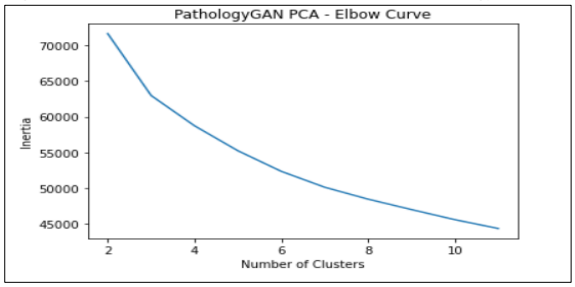
2.1 K-Means Clustering Algorithm

- An Iterative Algorithm that tries to partition the dataset into K distinct non-overlapping clusters where each data point belongs to only 1 cluster.
- K is number of clusters into which data points will be partitioned. Value of K is pre-defined.

2.1.1 Choosing Optimal Number Of Clusters : Elbow Method

- Inertia – Sum of the Squared distances of data points to their closest cluster center.
- Changed Number of clusters from 2 to 12 and calculated inertia.

Figure 1. Number of clusters vs Inertia for PathologyGAN PCA

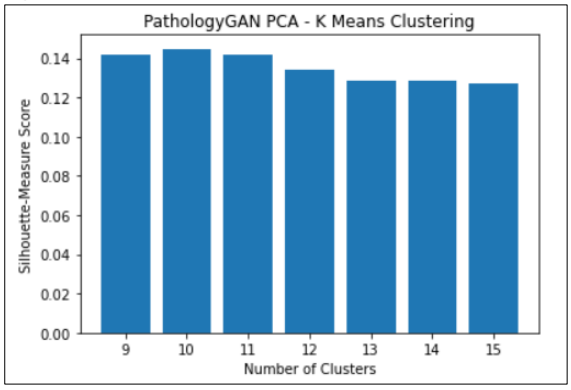


- Above graph shows Number of Clusters Vs calculated inertia for each number of clusters for PathologyGAN PCA dataset.
- Optimal Number of the clusters is the point after which inertia starts decreasing in linear fashion.
- As seen in the above graph, graph bends at the cluster value 3, so 3 is the Optimal number of clusters for PathologyGAN dataset.

2.1.2 Choosing Optimal Number Of Clusters : Silhouette Score

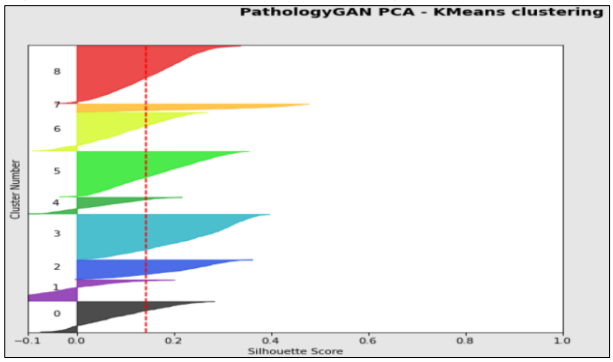
- Silhouette coefficient is metric used for calculating goodness of a clustering technique.
- Value ranges from -1 to 1.
- Value 1: Clusters are well apart from each other and clearly distinguished.
  Value 0: Clusters are indifferent, or the distance between clusters is not significant.
  Value -1: Clusters are assigned in the wrong way.
- Changed Number of clusters from 9 to 15 and calculated Silhouette Score for PathologyGAN PCA.

Figure 2. Number of clusters vs Silhouette Score for PathologyGAN PCA



- As seen in the above graph, Number of clusters 10 gives highest silhouette score which is close to 0.145.
- We already know that there are 9 tissue types and silhouette score with 9 clusters is close to 10 clusters.
- Below plot shows Silhoutte score for number of cluster values 9.

Figure 3. Silhouette Score for K-means algorithm with value of K (Number of clusters) as 9.

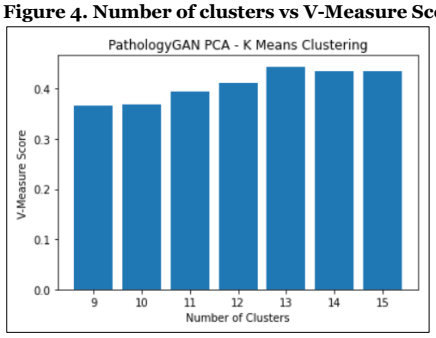


- Number of clusters 9 gives average Silhouette score close to 0.14.
- Number of clusters 9 is not an ideal choice for this dataset due to wide fluctuation of size of the Silhouette Plot.

2.1.3 Choosing Optimal Number Of Clusters : V-Measure Score

- V-Measure score is metric used for calculating goodness of a clustering technique.
- It calculates Homogeneity and Completeness of the clusters.
- Value ranges from 0 to 1.
- Value 0: Clusters are not homogenous and/or not complete
  Value 1: Clusters are homogenous and complete
- Changed Number of clusters from 9 to 15 and calculated V-Measure Score for PathologyGAN PCA.

Figure 4. Number of clusters vs V-Measure Score for PathologyGAN PCA

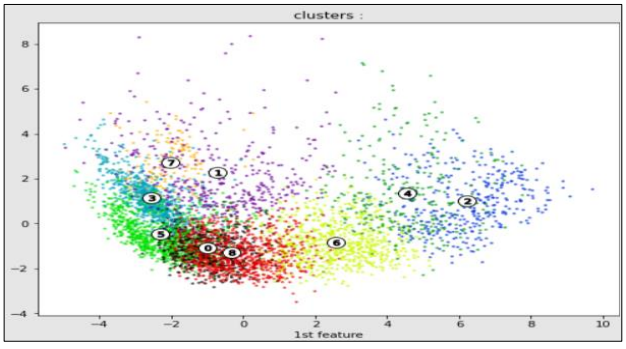


- As seen in the above graph, Number of clusters 13 gives highest silhouette score which is close to 0.44.
- We already know that there are 9 tissue types and V-measure score with 9 clusters is close to 10 clusters.

2.1.4 Taking K value (Number of Clusters) as 9 for K-Means

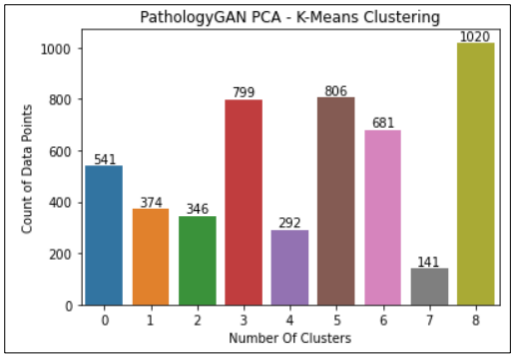
- We already know that there are 9 tissue types based on Task description.
- Plotting generated clusters with input number of clusters as 9 (K-Value 9) for K means algorithm.

Figure 5. Plot of 9 clusters with centroids for K-Means Algorithm



- As seen above, clusters are overlapped. So Number of Clusters as 9 is not an ideal choice for K-Mean clustering for PathologyGAN PCA dataset.
- Plotting Data points present in each cluster label.

Figure 6. K-Means Algorithm – Data Points in Each cluster - Representation – PathologyGAN



- As seen above, cluster 8 contains highest data points and cluster 7 contains the minimum data points.
- We have shown figures for K-Mean clustering algorithms for PathologyGAN PCA representation. We have applied similar methods for calculating Optimal Number of clusters for PathologyGAN – UMAP, VGG-16 PCA and VGG-16 UMAP representations.

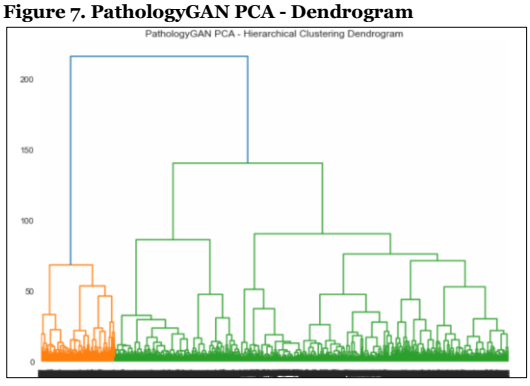
2.2 Hierarchical Clustering Algorithm

- An Iterative Algorithm that tries to build hierarchy of the clusters based on dissimilarities between the data points.
- There are 2 types of Hierarchical Clustering algorithm.
  - Agglomerative (Bottom-Up approach)
  - Divisive (Top-Down approach)
- We will use Agglomerative Hierarchical clustering algorithm for our dataset.
- Number of clusters parameter is optional in Hierachical clustering.
- If Number of clusters is not mentioned, Algorithm will calculate optimal number of clusters based on the data points.

2.2.1 Choosing Optimal Number Of Clusters : Dendrogram Method

- Tree Diagram showing relationship between similar data.

Figure 7. PathologyGAN PCA - Dendrogram



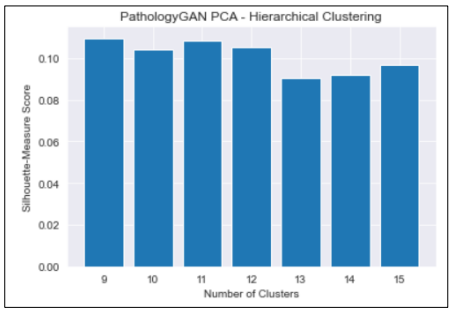
- We will keep the branches that have many observations and with a large distance above.
- Above graph shows that Optimal Number of Clusters for PathologyGAN PCA dataset is 3.

2.2.2 Choosing Optimal Number Of Clusters : Silhouette Score

- Silhouette coefficient is metric used for calculating goodness of a clustering technique.
- Value ranges from -1 to 1.

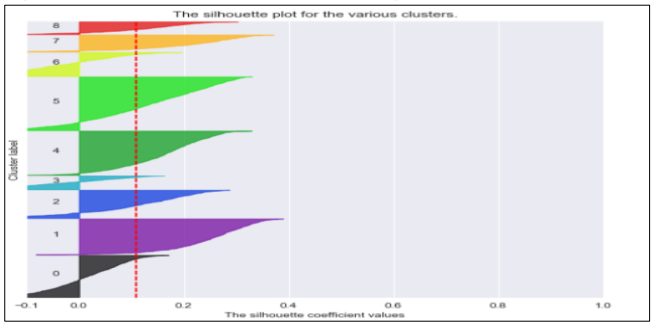
- Value 1: Clusters are well apart from each other and clearly distinguished.
  Value 0: Clusters are indifferent, or the distance between clusters is not significant.
  Value -1: Clusters are assigned in the wrong way.
- Changed Number of clusters from 9 to 15 and calculated Silhouette Score for PathologyGAN PCA.

Figure 8. Number of clusters vs Silhouette Score for PathologyGAN PCA



- As seen in the above graph, Number of clusters 9 gives highest silhouette score which is close to 0.109.
- We already know that there are 9 tissue types in our dataset.
- Below plot shows Silhoutte score for number of cluster values 9.

Figure 9. Silhouette Score for K-means algorithm with value of K (Number of clusters) as 9.

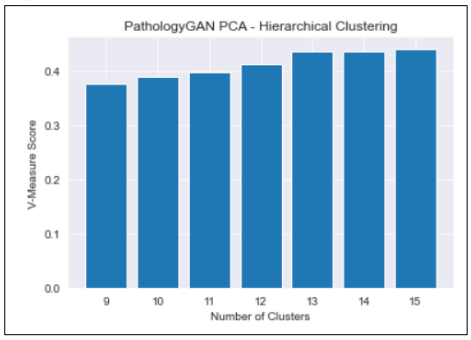


- Number of clusters 9 gives average Silhouette score close to 0.109.
- Number of clusters 9 is not an ideal choice for this dataset due to wide fluctuation of size of the Silhouette Plot.

2.2.3 Choosing Optimal Number Of Clusters : V-Measure Score

- V-Measure score is metric used for calculating goodness of a clustering technique.
- It calculates Homogeneity and Completeness of the clusters.
- Value ranges from 0 to 1.
- Value 0: Clusters are not homogenous and/or not complete
  Value 1: Clusters are homogenous and complete
- Changed Number of clusters from 9 to 15 and calculated V-Measure Score for PathologyGAN PCA.

Figure 11. Number of clusters vs V-Measure Score for PathologyGAN PCA

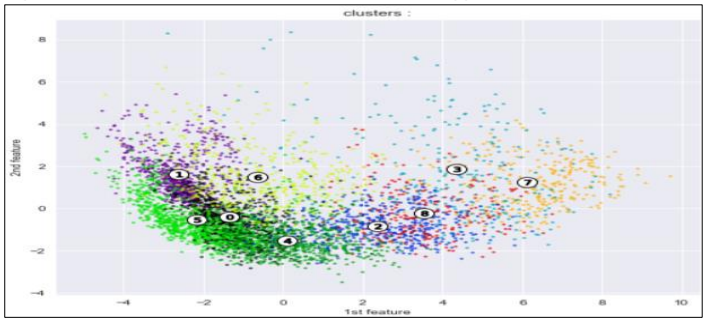


- As seen in the above graph, Number of clusters 15 gives highest silhouette score which is close to 0.45.
- We already know that there are 9 tissue types. So we will take number of clusters as 9.

2.1.4 Taking Number of Clusters as 9 for Hierarchical – Agglomerative Clustering

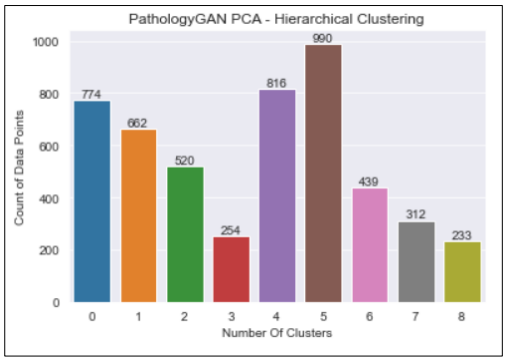
- We already know that there are 9 tissue types based on Task description.
- Plotting generated clusters with input number of clusters as 9 for Agglomerative – Hierarchical clustering algorithm.

Figure 12. Plot of 9 clusters with centroids for Agglomerative Hierarchical Clustering Algorithm



- As seen above, clusters are overlapped. So Number of Clusters as 9 is not an ideal choice for Agglomerative Hierarchical clustering for PathologyGAN PCA dataset.
- Plotting Data points present in each cluster label.

Figure 13. Hierarchical Algorithm – Data Points in Each cluster - Representation – PathologyGAN



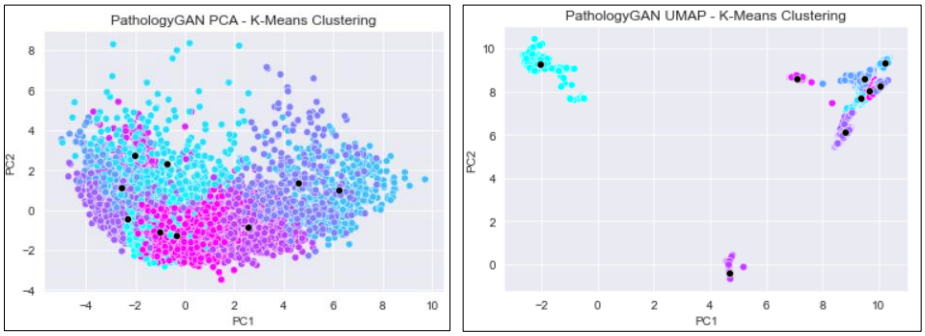
- As seen above, cluster 5 contains the maximum data points and cluster 8 contains the minimum data points.
- We have shown figures for Agglomerative Hierarchical clustering algorithms for PathologyGAN PCA representation. We have applied similar methods for calculating Optimal Number of clusters for PathologyGAN – UMAP, VGG-16 PCA and VGG-16 UMAP representations.

3 Results Discussion

3.1 K-Means and Hierarchical Clusters: Below plot shows clusters generated by K-Means and Agglomerative Hirarchical clustering algorithms for PathologyGAN and VGG16 representations.

Figure 14. K-Means Algorithm – Clusters and Centroid of clusters with value of K as 9 -

Representation - PathologyGAN



Representation – VGG16

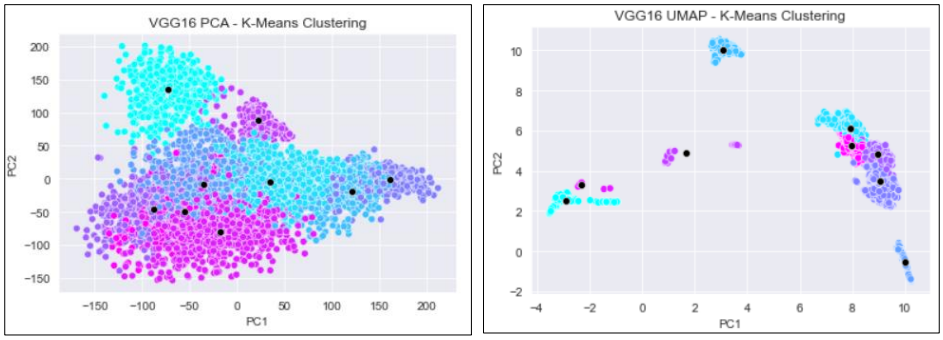
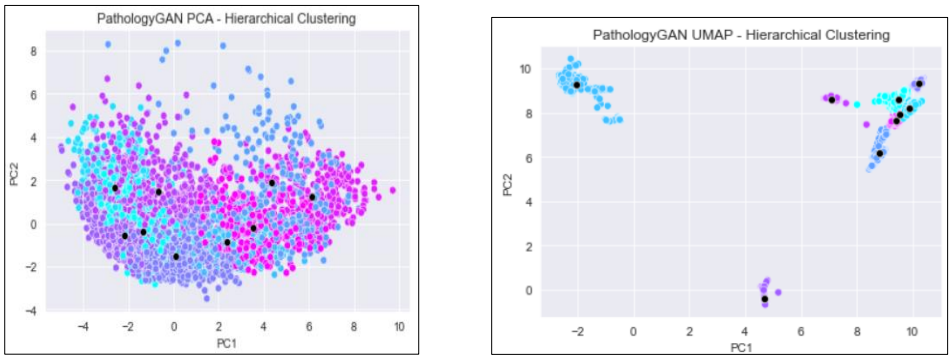
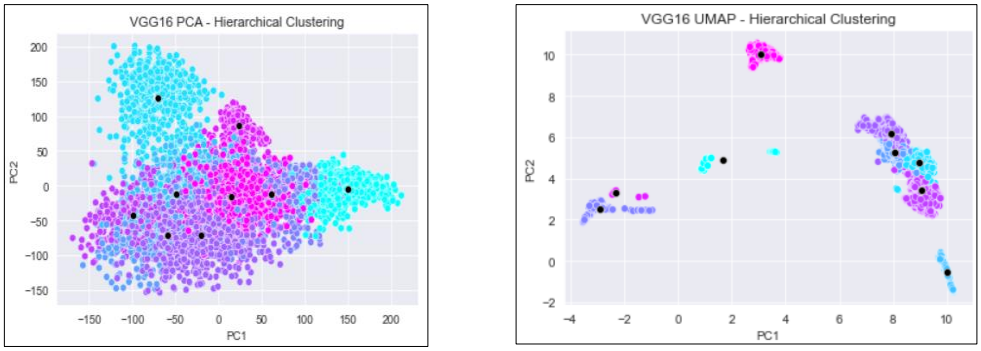


Figure 15. Agglomerative Hierarchical Clustering Algorithm – Clusters and Centroid of clusters with n_clusters as 9

Representation - PathologyGAN



Representation – VGG16

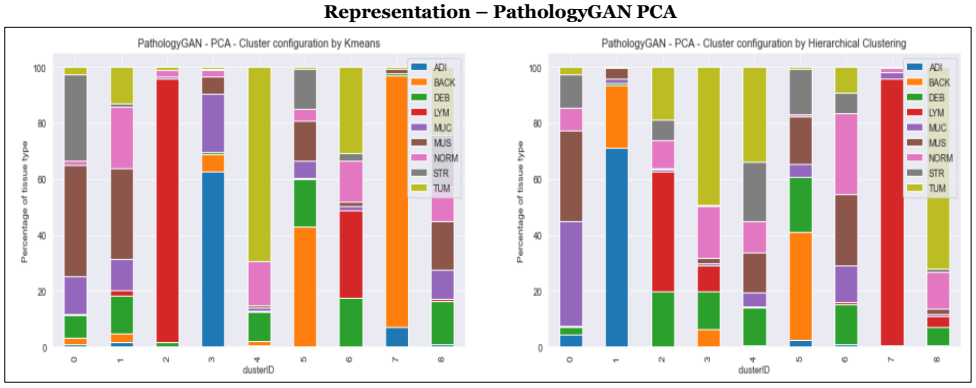


3.2 Percentage Of Tissue Types in Each Cluster

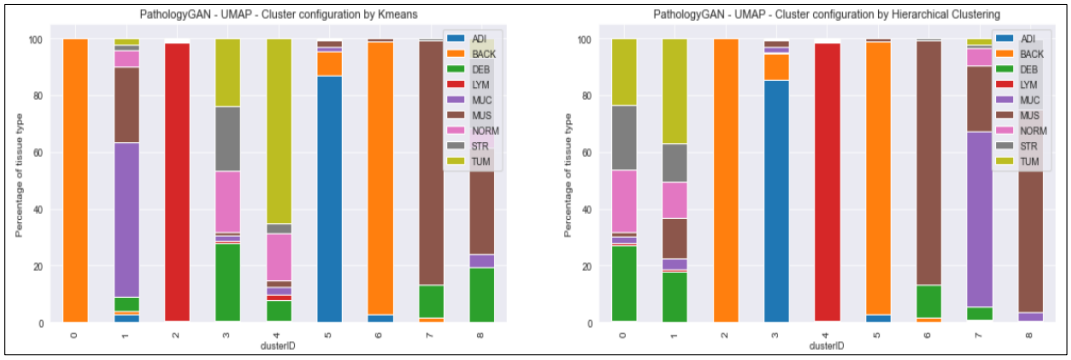
- Below plot shows percentage of different tissue types in clusters generated by K-Means and Hierarchical clustering algorithms for PathologyGAN and VGG16 representations.

Figure 16 percentage of tissue types in Clusters for Number of clusters 9 with K-Means and Hierarchical Clustering

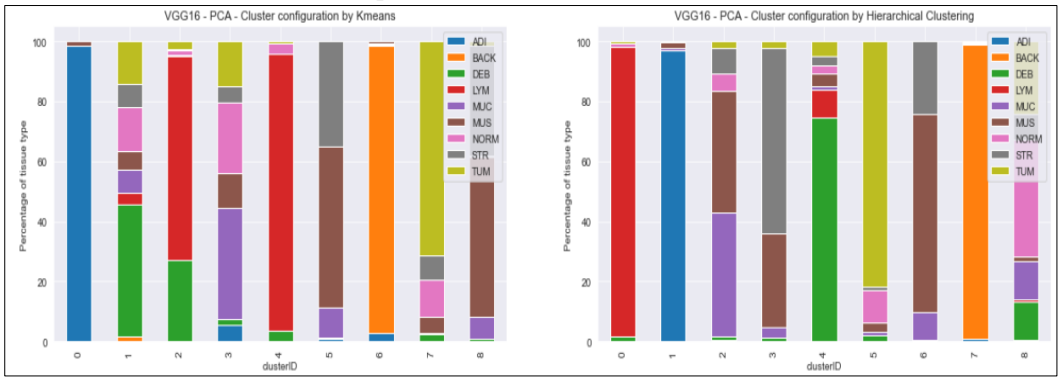
Representation – PathologyGAN PCA



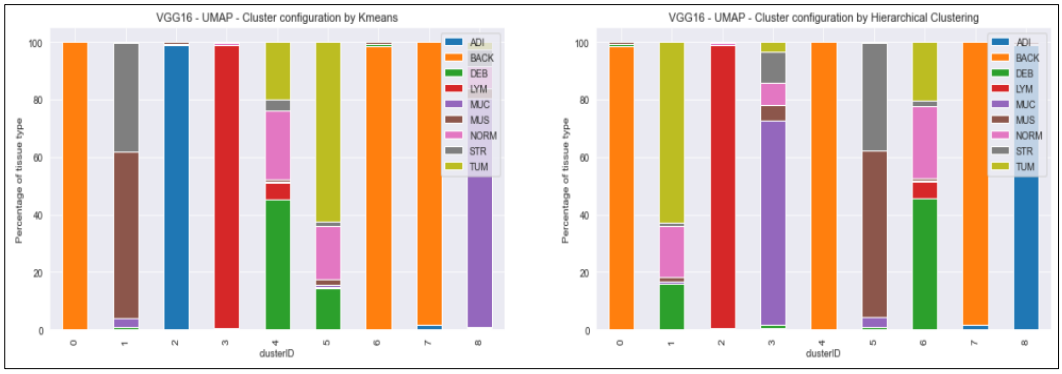
Representation – PathologyGAN UMAP



Representation – VGG16 PCA



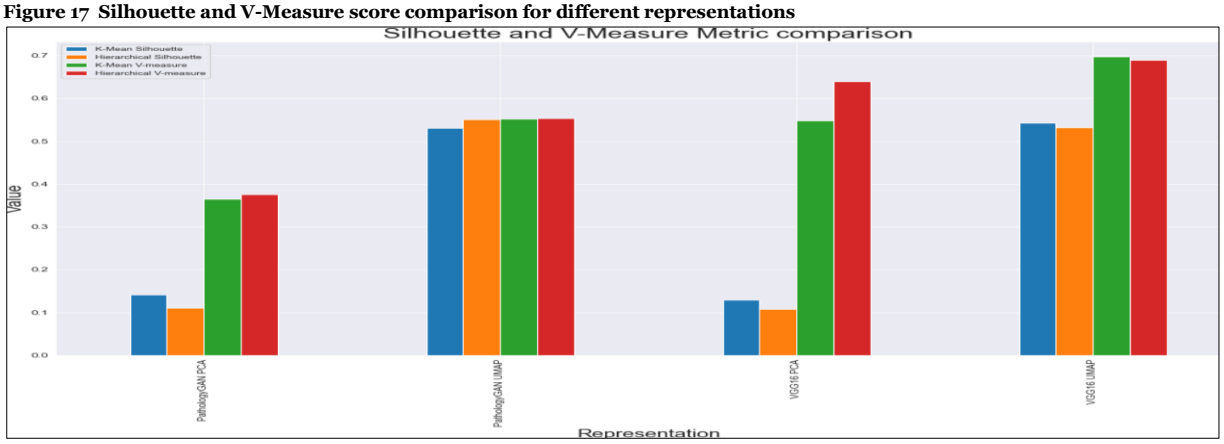
Representation – VGG16 UMAP



3.3 Cluster Performance Report (Comparison of Silhouette and V-Measure scores)

Representation	K-Mean Silhouette	Hierarchical Silhouette	K-Mean V-measure	Hierarchical V-measure
PathologyGAN PCA	0.141719	0.109597	0.365292	0.375037
PathologyGAN UMAP	0.530998	0.550981	0.551969	0.553202
VGG16 PCA	0.128872	0.107449	0.547975	0.639092
VGG16 UMAP	0.542247	0.532340	0.697235	0.689708

Figure 17 Silhouette and V-Measure score comparison for different representations



3.4 Observations

- By seeing above plots of clusters and comparison of Sillhouette and V-Measure score for K-Means clustering algorithm and Agglomerative Hierachical clustering algorithms, we can say that Sillhouette and V-Measure scores as well as clusters are better with UMAP than with PCA for both PathologyGAN and VGG16 representations.
- Silhoutte score is close to 0.1 with PCA representation and close to 0.5 with UMAP.
- K-means algorithm is giving slightly better performance for PathologyGAN PCA, VGG16 PCA and VGG16 UMAP. For PathologyGAN UMAP, Hierarchical algorithm is giving better scores compared to K-mean clustering algorithm.
- Silhoutte and V-measure scores for all 4 cases are low. Both the algorithms K-mean and Hierarchical are not giving good performance for these datasets as Silhoutte and V-measure score are not even close to 1.
- Clusters generated are overlapped and there is no clear separation between these clusters.

---

4 Conclusion

- Elbow method for K-mean clustering and Dendrogram for hierachical clustering were giving optimal number of clusters as 3 in case of PathologyGAN PCA and UMAP and VGG16 UMAP. For VGG16 PCA, both K-means Elbow and Hierarchical Dendrogram was giving ideal number of clusters as 6 but since we already knew that the dataset contains 9 tissue types, we checked the performance of algorithms with 9 clusters.
- UMAP representations for PathologyGAN and VGG16 datasets are giving better results compared to PCA representations.
- Clusters formed using K-mean and Hirarchical algorithms are not very inseparable and they are some showing outliers as well.
- For both algorithms K-mean and Hierarchical-Agglomerative, we specified number of clusters as 9 while training the algorithm. If we would have used algorithms like Louvain or GMM which interprets optimal number of clusters from data itself, we may have got the better results.
- Metrics values from both K-mean and Hierarchical - Agglomerative algorithm are very low. So we cannot completely rely only on these 2 models for our datatsets. We may need to try with some other clustering algorithm and compare the performance to check if they are giving better results.