Clustering Document

Fatemeh Dehbashi, Mina Afzali

April 12, 2024

Contents

1 Introduction

In this project, our aim is to employ various clustering algorithms to group images of individuals. Our database encompasses a variety of photos featuring 15 distinct individuals. The objective is to develop and execute a system that effectively segregates these individuals' images from one another.

2 Phase 1: Feature Extracting

In this phase, our approach began with feature extraction from the images, generating a unique feature vector for each. To achieve this, we utilized two models: VGG16 and ResNet50:

2.1 The VGG16 Model

- VGG16 is a deep convolutional neural network model used for image classification tasks. The network is composed of 16 layers of artificial neurons, which each work to process image information incrementally and improve the accuracy of its predictions.

- VGG16 is used for image recognition and classification in new images. The pre-trained version of the VGG16 network is trained on over one million images from the ImageNet visual database, and is able to classify images into 1,000 different categories with 92.7 percent top-5 test accuracy. VGG16 can be applied to determine whether an image contains certain items, animals, plants and more.

2.2 The ResNet50 Model

- ResNet-50 is a 50-layer convolutional neural network (48 convolutional layers, one MaxPool layer, and one average pool layer). Residual neural networks are a type of artificial neural network (ANN) that forms networks by stacking residual blocks.

- The ResNet architecture follows two basic design rules. First, the number of filters in each layer is the same depending on the size of the output feature map. Second, if the feature map's size is halved, it has double the number of filters to maintain the time complexity of each layer.

3 Phase 2: Dimensionality Reduction

In Machine Learning and Statistics,in order to build a good performing model we try to pass on those features in the dataset that are significant to one another. Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.

3.1 Principal Component analysis (PCA)

PCA is an unsupervised linear dimensionality reduction and data visualization technique for very high dimensional data. As having high dimensional data is very hard to gain insights from adding to that, it is very computationally intensive. The main idea behind this technique is to reduce the dimensionality of data that is highly correlated by transforming the original set of vectors to a new set which is known as Principal component. PCA tries to preserve the Global Structure of data i.e when converting d-dimensional data to d'-dimensional data then it tries to map all the clusters as a whole due to which local structures might get lost. Application of this technique includes Noise filtering, feature extractions, stock market predictions, and gene data analysis.

3.2 t-distributed stochastic neighbourhood embedding (t-SNE)

t-SNE is also a unsupervised non-linear dimensionality reduction and data visualization technique. The math behind t-SNE is quite complex but the idea is simple. It embeds the points from a higher dimension to a lower dimension trying to preserve the neighborhood of that point. Unlike PCA it tries to preserve the Local structure of data by minimizing the Kullback–Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map. This technique finds application in computer security research, music analysis, cancer research, bioinformatics, and biomedical signal processing.

4 Phase 3: Clustering Algorithms

4.1 K-Means

K-Means is a popular clustering algorithm that partitions data into K clusters. It works by iteratively assigning each data point to the nearest cluster centroid and then recalculating the centroids based on the mean of the points assigned to each cluster. It aims to minimize the within-cluster variance, resulting in tight clusters around centroids.

4.2 Mean Shift

MeanShift is a non-parametric clustering algorithm that doesn't require specifying the number of clusters beforehand. It works by iteratively shifting data points towards the mode of the underlying data distribution until convergence. The bandwidth parameter determines the size of the region for which to estimate the density around each data point.

4.3 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering algorithm that identifies clusters as regions of high density separated by regions of low density. It doesn't require specifying the number of clusters in advance and is capable of identifying clusters of arbitrary shapes. It classifies points as core, border, or noise based on their density relative to the specified parameters: epsilon (eps) and minimum number of points (min_samples).

4.4 Agglomerative Clustering

Agglomerative clustering is a hierarchical clustering method that recursively merges the nearest pairs of clusters until only a specified number of clusters remain. The linkage parameter determines the criteria for merging clusters, with options including **"average"**, **"single"**, and **"complete"** linkage. "Average" linkage merges clusters based on the average distance between their points, "single" linkage merges based on the minimum distance, and "complete" linkage merges based on the maximum distance.

5 Phase 4: Results

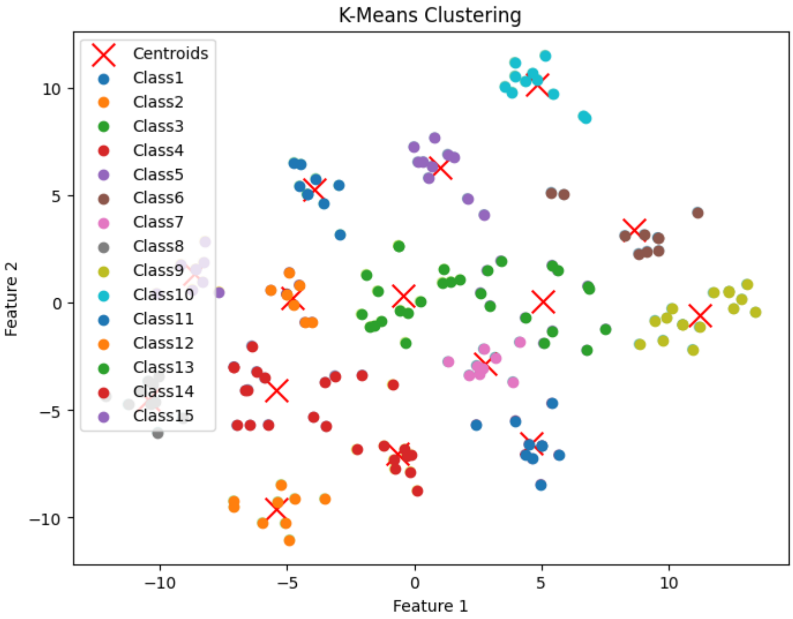
DBSCAN and K-Means are among two best clustering algorithms that we used.

5.1 K-Means

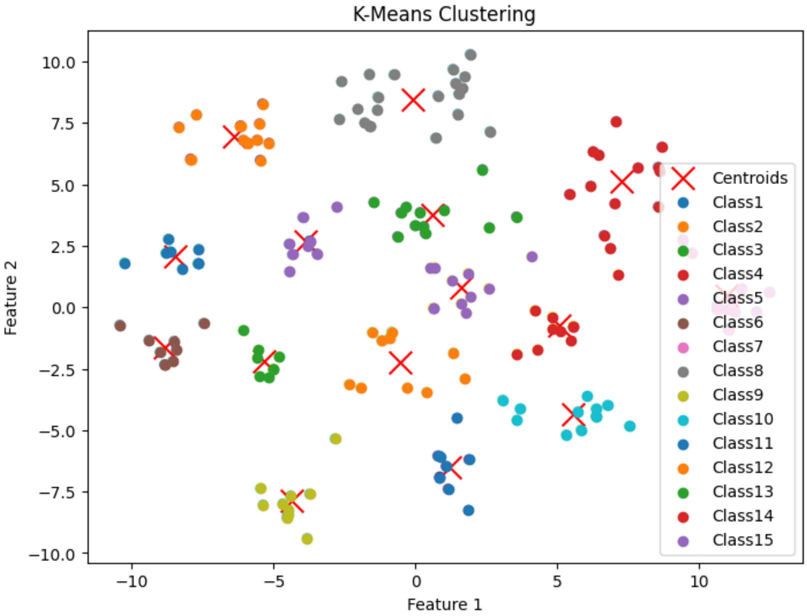
- Advantages:
  - Easy to implement and interpret
  - Fast and efficient for large datasets
  - Works well with spherical clusters
- Disadvantages:
  - Sensitive to initial centroid positions
  - Can converge to local optima

5.1.1 Plots

- ResNet50 Model:

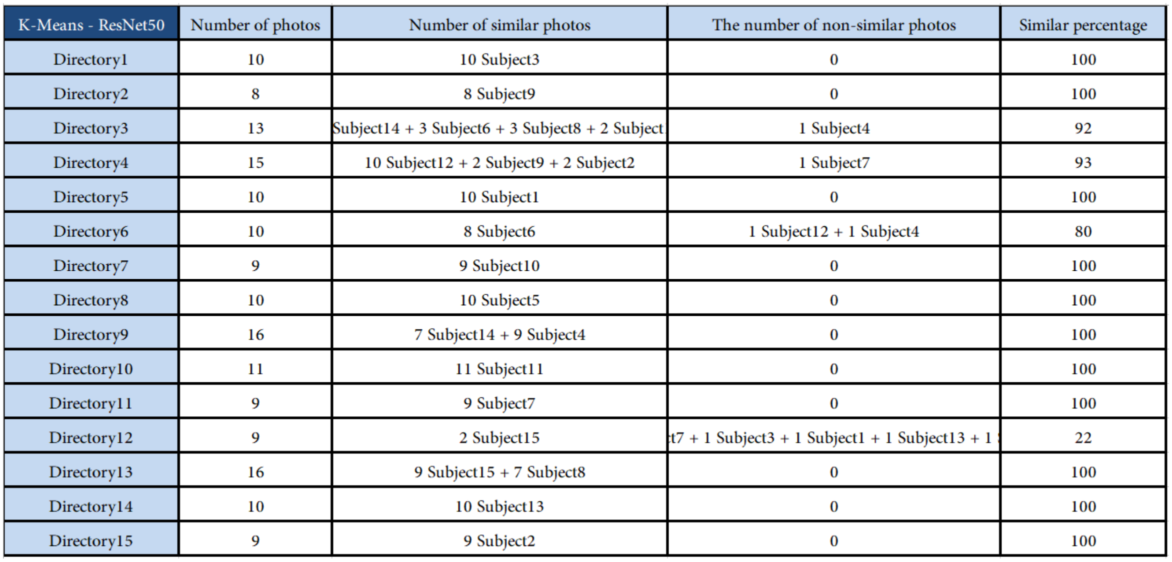


- VGG16 Model:



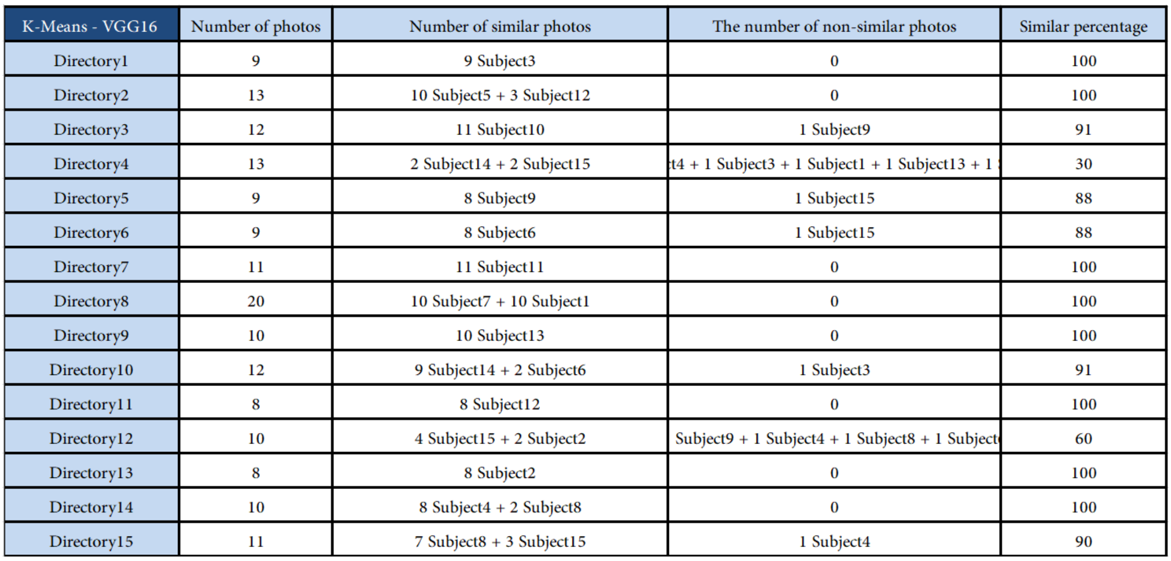
5.1.2 Results

ResNet50 Model:



K-Means - ResNet50	Number of photos	Number of similar photos	The number of non-similar photos	Similar percentage
Directory1	10	10 Subject3	0	100
Directory2	8	8 Subject9	0	100
Directory3	13	Subject14 + 3 Subject6 + 3 Subject8 + 2 Subject	1 Subject4	92
Directory4	15	10 Subject12 + 2 Subject9 + 2 Subject2	1 Subject7	93
Directory5	10	10 Subject1	0	100
Directory6	10	8 Subject6	1 Subject12 + 1 Subject4	80
Directory7	9	9 Subject10	0	100
Directory8	10	10 Subject5	0	100
Directory9	16	7 Subject14 + 9 Subject4	0	100
Directory10	11	11 Subject11	0	100
Directory11	9	9 Subject7	0	100
Directory12	9	2 Subject15	t7 + 1 Subject3 + 1 Subject1 + 1 Subject13 + 1	22
Directory13	16	9 Subject15 + 7 Subject8	0	100
Directory14	10	10 Subject13	0	100
Directory15	9	9 Subject2	0	100

VGG16 Model:



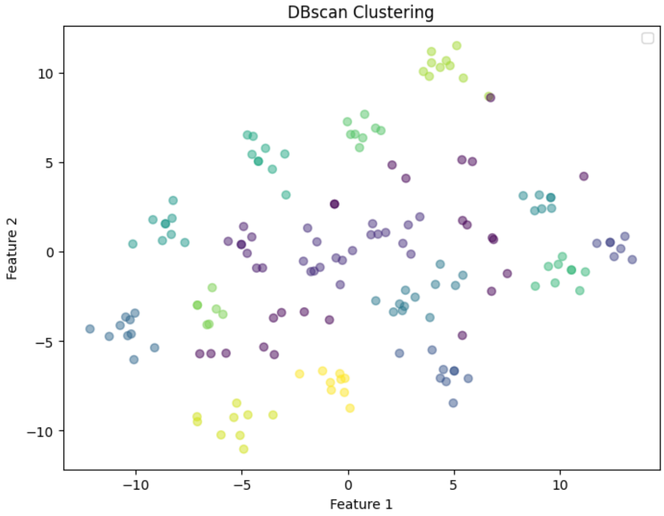
K-Means - VGG16	Number of photos	Number of similar photos	The number of non-similar photos	Similar percentage
Directory1	9	9 Subject3	0	100
Directory2	13	10 Subject5 + 3 Subject12	0	100
Directory3	12	11 Subject10	1 Subject9	91
Directory4	13	2 Subject14 + 2 Subject15	t4 + 1 Subject3 + 1 Subject1 + 1 Subject13 + 1	30
Directory5	9	8 Subject9	1 Subject15	88
Directory6	9	8 Subject6	1 Subject15	88
Directory7	11	11 Subject11	0	100
Directory8	20	10 Subject7 + 10 Subject1	0	100
Directory9	10	10 Subject13	0	100
Directory10	12	9 Subject14 + 2 Subject6	1 Subject3	91
Directory11	8	8 Subject12	0	100
Directory12	10	4 Subject15 + 2 Subject2	Subject9 + 1 Subject4 + 1 Subject8 + 1 Subject	60
Directory13	8	8 Subject2	0	100
Directory14	10	8 Subject4 + 2 Subject8	0	100
Directory15	11	7 Subject8 + 3 Subject15	1 Subject4	90

5.2 DBSCAN

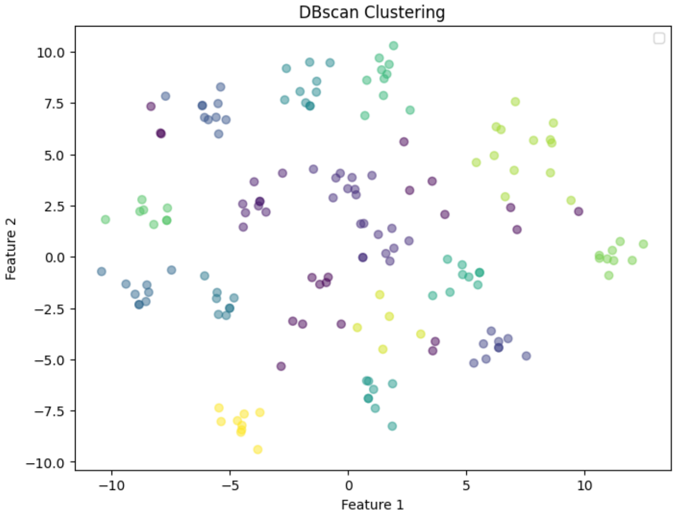
One of the advantages of DBSCAN is that it can find clusters of arbitrary shapes and sizes, unlike K-Means which assumes spherical clusters. DBSCAN is also robust to noise and outliers since they are not assigned to any cluster. However, DBSCAN can be sensitive to the choice of distance metric and parameters such as the radius and minimum number of points required to form a cluster.

5.2.1 Plots

- ResNet50 Model:

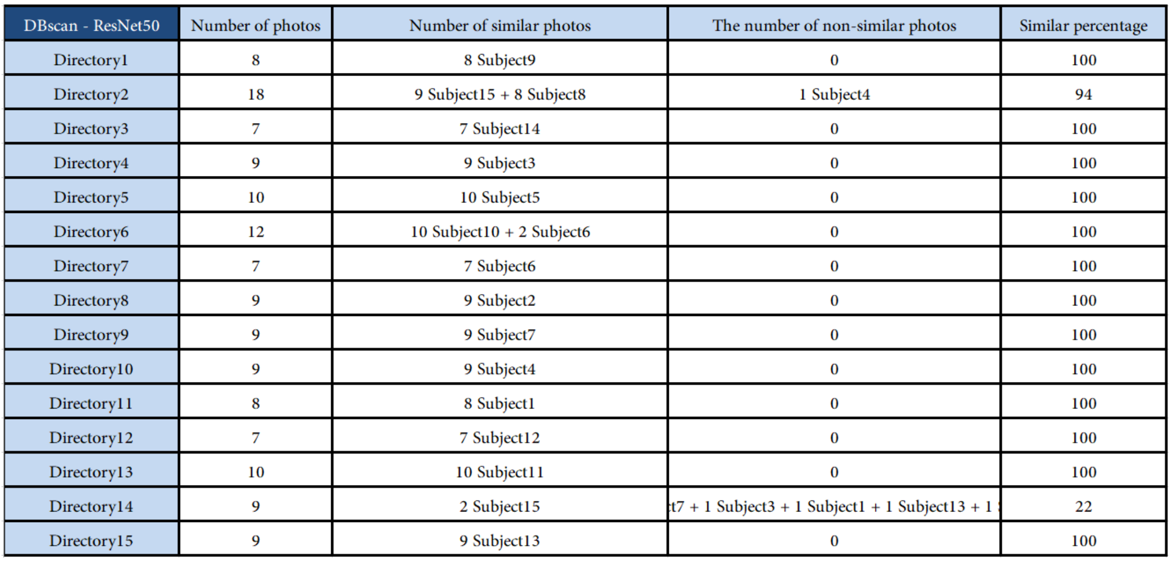


- VGG16 Model:



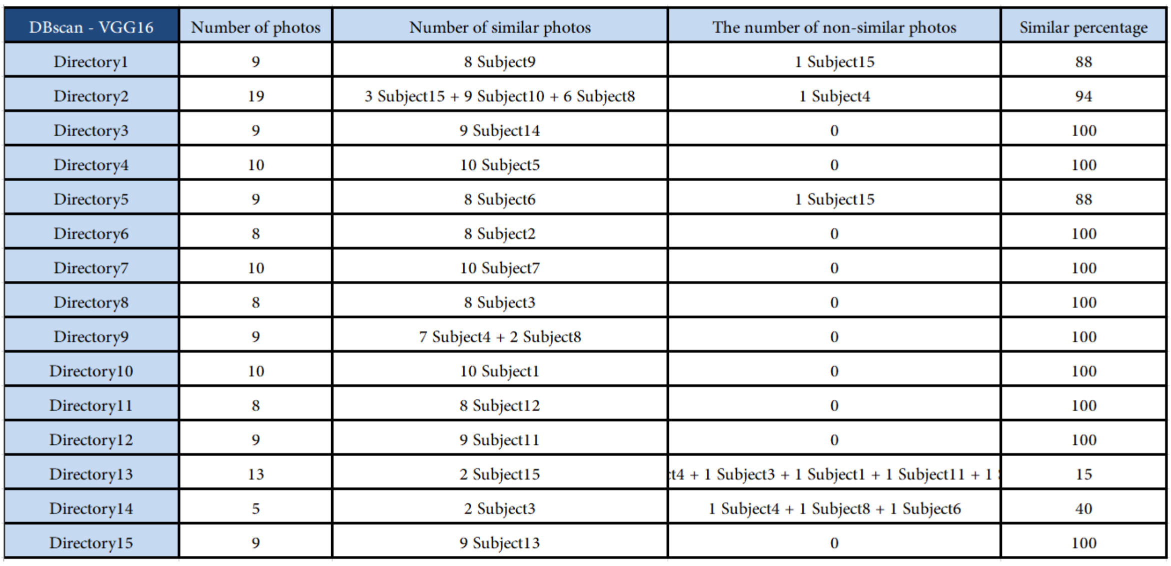
5.2.2 Results

ResNet50 Model:



DBscan - ResNet50	Number of photos	Number of similar photos	The number of non-similar photos	Similar percentage
Directory1	8	8 Subject9	0	100
Directory2	18	9 Subject15 + 8 Subject8	1 Subject4	94
Directory3	7	7 Subject14	0	100
Directory4	9	9 Subject3	0	100
Directory5	10	10 Subject5	0	100
Directory6	12	10 Subject10 + 2 Subject6	0	100
Directory7	7	7 Subject6	0	100
Directory8	9	9 Subject2	0	100
Directory9	9	9 Subject7	0	100
Directory10	9	9 Subject4	0	100
Directory11	8	8 Subject1	0	100
Directory12	7	7 Subject12	0	100
Directory13	10	10 Subject11	0	100
Directory14	9	2 Subject15	t7 + 1 Subject3 + 1 Subject1 + 1 Subject13 + 1	22
Directory15	9	9 Subject13	0	100

VGG16 Model:



DBscan - VGG16	Number of photos	Number of similar photos	The number of non-similar photos	Similar percentage
Directory1	9	8 Subject9	1 Subject15	88
Directory2	19	3 Subject15 + 9 Subject10 + 6 Subject8	1 Subject4	94
Directory3	9	9 Subject14	0	100
Directory4	10	10 Subject5	0	100
Directory5	9	8 Subject6	1 Subject15	88
Directory6	8	8 Subject2	0	100
Directory7	10	10 Subject7	0	100
Directory8	8	8 Subject3	0	100
Directory9	9	7 Subject4 + 2 Subject8	0	100
Directory10	10	10 Subject1	0	100
Directory11	8	8 Subject12	0	100
Directory12	9	9 Subject11	0	100
Directory13	13	2 Subject15	t4 + 1 Subject3 + 1 Subject1 + 1 Subject11 + 1	15
Directory14	5	2 Subject3	1 Subject4 + 1 Subject8 + 1 Subject6	40
Directory15	9	9 Subject13	0	100

5.3 Comparing DBSCAN and K-Means

DBSCAN and K-Means are among two best clustering algorithms that we used.

- Differences between the two algorithms:
- DBSCAN is a density-based clustering algorithm, whereas K-Means is a centroid-based clustering algorithm.
- DBSCAN can discover clusters of arbitrary shapes, whereas K-Means assumes that the clusters are spherical.
- DBSCAN does not require the number of clusters to be specified in advance, whereas K-Means requires the number of clusters to be specified.
- DBSCAN is less sensitive to initialization than K-Means.

DBscan is better than K-means in clustering in this particular case because DBscan is able to handle noisy data points and outliers effectively. As seen in the results, DBscan has a lower number of noise points compared to K-means, which means DBscan is more robust in dealing with outliers. Additionally, DBscan does not require specifying the number of clusters beforehand, unlike K-means which requires the number of clusters to be set manually. This flexibility allows DBscan to adapt better to the data and find more accurate clusters, resulting in a higher Rand Index score.

6 Phase : Rand-Index

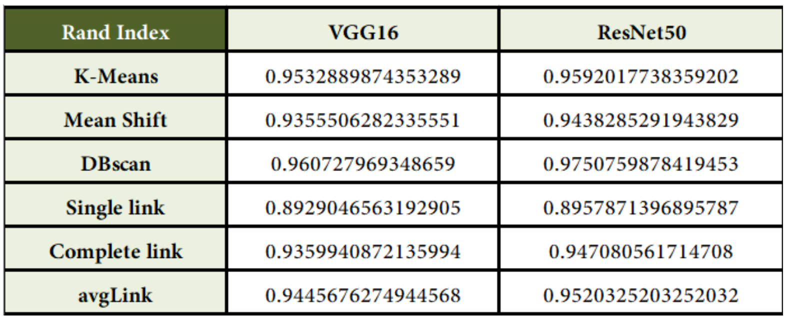
The Rand Index is a measure of similarity between two data clusterings, where TP (true positives) is the number of pairs of elements that are in the same cluster in both the true and predicted clusterings, TN (true negatives) is the number of pairs of elements that are in different clusters in both clusterings, FP (false positives) is the number of pairs of elements that are in the same cluster in the predicted clustering, but not in the true clustering, and FN (false negatives) is the number of pairs of elements that are in different clusters in the predicted clustering, but in the same cluster in the true clustering.

$$RandIndex = \frac{TP + TN}{TP + TN + FP + FN}$$

While the Rand Index is commonly used to evaluate clustering algorithms, it is not without its limitations. One reason why the Rand Index may not work well in some cases is that it does not take into account the actual clustering structure or the size of the clusters. For example, if a clustering algorithm produces clusters that are very different in size or are highly overlapping, the Rand Index may still result in a high value, even though the clustering is not good.

Additionally, the Rand Index can be influenced by the number of data points or clusters in the dataset. In datasets with a large number of data points or clusters, the Rand Index may tend to be higher even if the clustering is not accurate.

In summary, the Rand Index can be a useful measure of similarity between two clusterings, but it is important to consider its limitations and potentially use other evaluation metrics or visualizations to get a more complete understanding of the clustering quality.



Rand Index	VGG16	ResNet50
K-Means	0.9532889874353289	0.9592017738359202
Mean Shift	0.9355506282335551	0.9438285291943829
DBscan	0.960727969348659	0.9750759878419453
Single link	0.8929046563192905	0.8957871396895787
Complete link	0.9359940872135994	0.947080561714708
avgLink	0.9445676274944568	0.9520325203252032