



Comparative Efficiency of DBSCAN, K-means, and Their Combination in Image Processing

**Elizabeth Odrick Koola, Charles Okanda Nyatega, Joseph Sospeter Salawa, Juma Said Ally, Cuthbert John Karawa, Phocas Sebastian, and Richard Mwanjalila*
Department of Electronics and Telecommunication Engineering,
Mbeya University of Science and Technology (MUST)

***Corresponding Email:** koolaeliza@gmail.com

Abstract

Clustering techniques are vital for image analysis applications like object detection, pattern recognition, and image segmentation. The research evaluates effectiveness and efficiency of three clustering techniques K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and a new hybrid model that combines the two are compared in this research in the context of analyzing MRI image dataset. Finding a technique that performs well across various kinds of image analysis applications in terms of computing efficiency and clustering quality is the main goal. In this study, a hybrid model is presented to overcome the shortcomings of individual techniques. The hybrid model aims at achieving a balanced performance that increases computational efficiency and clustering accuracy in MRI image analysis by combining the characteristics of DB-SCAN and K-means. This methodology makes use of DBSCAN's capacity to detect clusters with complicated shapes and densities combined with K-means' effectiveness in handling huge amount of data with simple cluster structures, the research significance is on how it potentially changes how the information is extracted from MRI images using the hybrid technique also it advances the field of medical image analysis. Silhouette Score is an Evaluation metric employed to assess clustering quality, and execution time metric as Computational efficiency measure. Based to our research, the hybrid model achieves a balanced performance on analyzing variety images in dataset by utilizing the advantages of both K-means and DBSCAN with the aim of opening the door to more precise and effective analysis in the field of medical imaging as well as beyond.

Keywords: Clustering, DBSCAN, K-Means, Hybrid, MRI Images

INTRODUCTION

As explained by *Magnetic Resonance Imaging (MRI)* (n.d.), Nyatega, Qiang, Adamu, and Kawuwa (2022) and Nyatega, Qiang, Adamu, Younis, and Kawuwa (2021) In modern technology, Magnetic Resonance Imaging (MRI), is essential for the diagnosis and comprehension of a wide range of medical disorders. Because of the amount of information present in MRI pictures, complex analytical techniques are required to extract significant insights. A machine learning technique called cluster analysis or clustering is an unsupervised learning approach and a widely used statistical data analysis method across numerous domains that involves organizing sets of items so that the objects in the same group, or cluster, are more similar to each other than to the objects in other groups *Cluster Analysis* (2019). As claimed by a (n.d.) this unsupervised machine learning technique is important in identifying patterns and features that a particular group of data possesses, so it is important in identifying what is inside MRI images during the analysis process. The research, "Comparative Efficiency of DBSCAN, K-means, and Their Combination in Image Processing" seeks to improve the precision and



effectiveness of MRI image processing by investigating and utilizing the potential of combined clustering techniques. Since it is often challenging for traditional image analysis techniques to capture the smallest details and complexity found in medical images. MRI scans are diverse and complex, still requiring a more advanced approach. Therefore, the study uses a collaborative approach of combining two different clustering approaches (K-Means and DBSCAN (Density-Based Spatial Clustering of Applications with Noise)) to provide a more detailed and sophisticated interpretation of the MRI data. The work of (Boulis & Ostendorf, 2004) and (Khedairia & Khadir, 2022) inspire this effort by illustrating how integrated clustering approaches might enhance the accuracy of medical image segmentation. Building on this framework, the research work aims to make a significant contribution to the area by offering a method for MRI image analysis that is more detailed and complex. We hope that by investigating using mixed clustering techniques, we will advance the field of medical image analysis and help achieve the larger objective of better patient outcomes via more precise and effective image analysis applications.

RELATED WORKS

Segmentation (clustering) Algorithms and Techniques in MRI Analysis

Image segmentation is one of the applications of clustering image analysis, Despotović, Goossens, and Philips (2015) and Guan, Zhao, Nyatega, and Li (2023) explain image segmentation as one of the most essential responsibilities in medical image analysis, which often serves as the initial and most important stage in many clinical applications. In brain magnetic resonance imaging analysis, image segmentation is commonly used to determine and visualize the brain's anatomical structures as well as to investigate changes in the brain caused by different diseases. According to (*DBSCAN vs. K-Means: A Guide in Python*, n.d.), To find groups of similar items in a dataset, clustering is an useful unsupervised machine learning technique. According to Velunachiyar and Sivakumar (2023), Clustering techniques can be distinguished as partitioning-based clustering, hierarchical-based clustering, grid-based clustering, model-based clustering, and fuzzy-based clustering that have been widely employed for their ability to group similar regions within MRI images. Each algorithm brings unique strengths and limitations, prompting researchers to explore different new clustering techniques and the integration of multiple clustering methods to enhance the robustness of segmentation results. Figure 1 below shows how clustering techniques are distinguished.

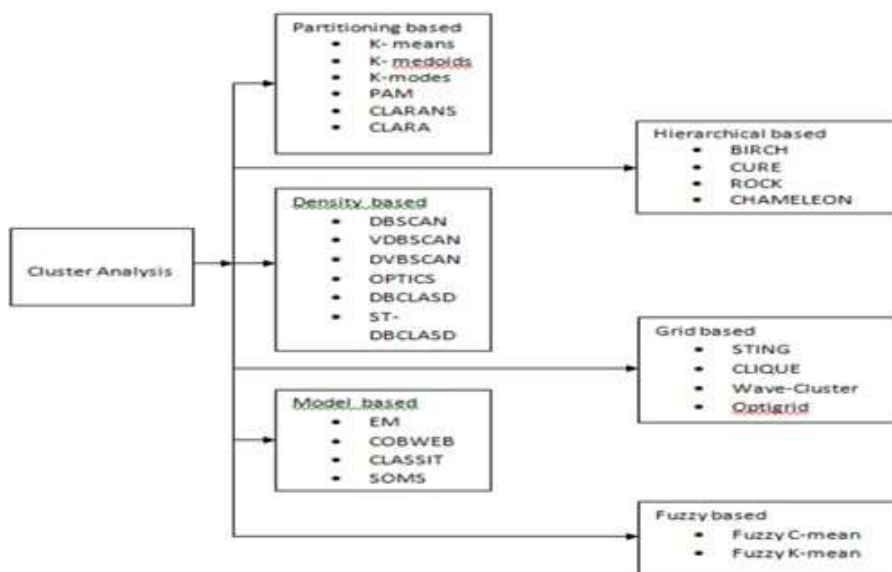


Figure 1: Distinguished Clustering Techniques (Velunachiyar & Sivakumar, 2023)

K-Means in Image Processing

According to *Introduction to Image Segmentation with K-Means clustering* (n.d.), The area of interest can be separated from the rest of the image using the unsupervised K-Means clustering technique. The provided data is divided or clustered into K-clusters or sections based on the K-centroids. When dealing with unlabeled data i.e., data without precisely defined categories or groups the algorithm is employed. Finding particular groups based on data similarity, with K being the number of groups, and minimizing the sum of squared distances between all points and the cluster center is the purpose. The below figure 2 is a K-means clustering example of Customers of a shopping mall that have been grouped into 5 clusters based on their income and spending score. With yellow dots presenting the Centroid of each cluster.

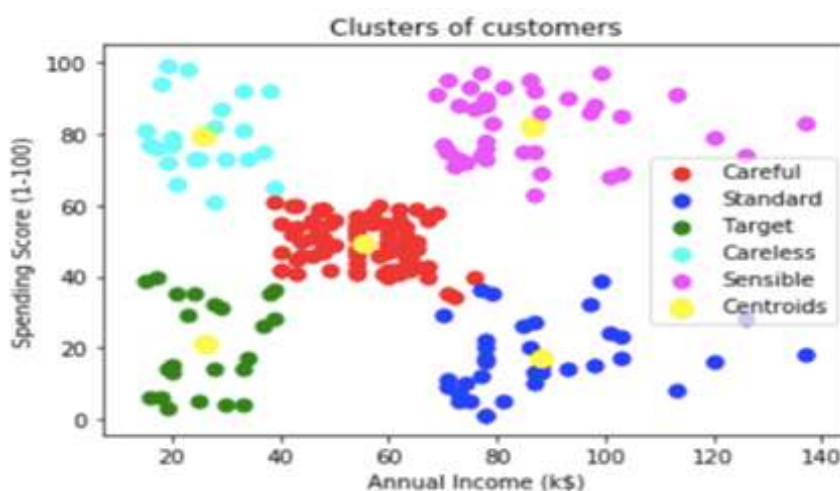




Figure 2: K-Means Clustering Example (*K-means Clustering Algorithm: Applications, Types, and Demos [Updated]* | Simplilearn, n.d.)

K-means Clustering Algorithm: Applications, Types, and Demos [Updated] Simplilearn (n.d.) claimed that K-means clustering has become one of the most widely used clustering algorithms, where K is the number of clusters and the algorithm works by following the below steps.

Select the K number of clusters that you wish to examine.

Assign the data points at randomness to any one of the k clusters.

Next, identify each cluster's center.

Evaluate the separation between each cluster's center and the data points.

Reassign the data points to the closest clusters based on each individual data point's distance from the cluster.

Determine the new cluster center once more.

Repeat steps 4,5 and 6 until the specified amount of iterations is reached or until the data points no longer cause the clusters to change.

K-means Clustering Algorithm: Applications, Types, and Demos [Updated] Simplilearn (n.d.) perform K-Means clustering on a butterfly image and concluded that Since the K-means algorithm can identify more classes or clusters of colors, the image becomes sharper and more distinct as the value of k increases. Furthermore, the algorithm performs better while working with less data sets, as it can correctly split objects in images.

DBSCAN in Image processing

According to Ester, Kriegel, Sander, and Xu (1996), The density-based approach to clusters used by DBSCAN is meant to find clusters of any shape. It requires just one input parameter, and it assists the user choose the right value. *DBSCAN vs. K-Means: A Guide in Python* (n.d.) claimed that because of the way how clusters are constructed in DBSCAN algorithm many problems that can arise when applying K-means clustering on an image are mitigated, one wonderful thing about DBSCAN is that it doesn't require a minimum or maximum number of clusters to be used. To compute the distance between values, all you need is a function and some parameters for what defines a "close" distance. Additionally, DBSCAN yields more realistic outcomes than k-means for an extensive number of densities. The below figure 3 shows this fact.



Figure 3: DBSCAN VS K-Means (*DBSCAN vs. K-Means: A Guide in Python*, n.d.)



Comparative Studies of DBSCAN and K-means

According to *DBSCAN vs. K-Means: A Guide in Python* (n.d.), The distance between centroids and points is the basis for the K-means clustering algorithm, which works by dividing data into finite

clusters, DBSCAN is a density-based clustering algorithm that puts data points together based on their closeness to one another, Algorithm selection depends on problem and characteristics of datasets because each algorithm has got their advantages and disadvantages. Sharma (2020) performed K-means and DBSCAN algorithms on the same Iris Dataset using Scikit Learn and obtained different results where the DBSCAN algorithm produces clusters that are clearly defined compared to those formed by the K-Means algorithm and concluded that when to use which of these two clustering techniques, depends on the problem you want to solve. Even though K-Means is the most popular clustering technique compared to DBSCAN, there are cases where, using DBSCAN results in better-defined clusters. So, it is important to identify the nature of your case's datasets before selecting a clustering technique.

Clustering Algorithm Combining Approaches

The research aims at combining K-means and DBSCAN clustering approaches in model generation. According to Jafarzadegan, Safi-Esfahani, and Beheshti (2019) Finding the right approach for a given problem can be challenging since applying one approach to multiple problems yields inconsistent outcomes. Suggestions for combinational approaches have been made to address this issue. In order to solve the high-dimensional data clustering issue, (Yan & Liu, 2022) explained an integrated clustering method for high-dimensional data since the traditional clustering algorithms, like K-means, do not perform well on high-dimensional data due to the multiplicity of irrelevant attributes, limited distribution, and complex calculations. Strehl and Ghosh (n.d.) identified several application scenarios for the resultant knowledge reuse framework named cluster ensembles. Three different qualitative application scenarios were evaluated. The 1st is where the original clusters were formed based on non-identical sets of features, 2nd is where the original clustering algorithms worked on non-identical sets of objects, and 3rd scenario is where a common data set is used. All these scenarios share the same main purpose which is to improve the quality and robustness of the solution. Jain (2012) performed a hybrid clustering approach on K-mean and K-harmonic mean (KHM) clustering algorithms and tested the proposed algorithm with five different datasets and it was found the combined algorithm robustness is much better compared to individual traditional algorithms. Pedro (2022) uses one of the clustering combination approaches called ensemble clustering where he concludes that ensemble K-means will make it possible to find much more complex shapes just like HDBSCAN.

RESULTS AND DISCUSSION

Results

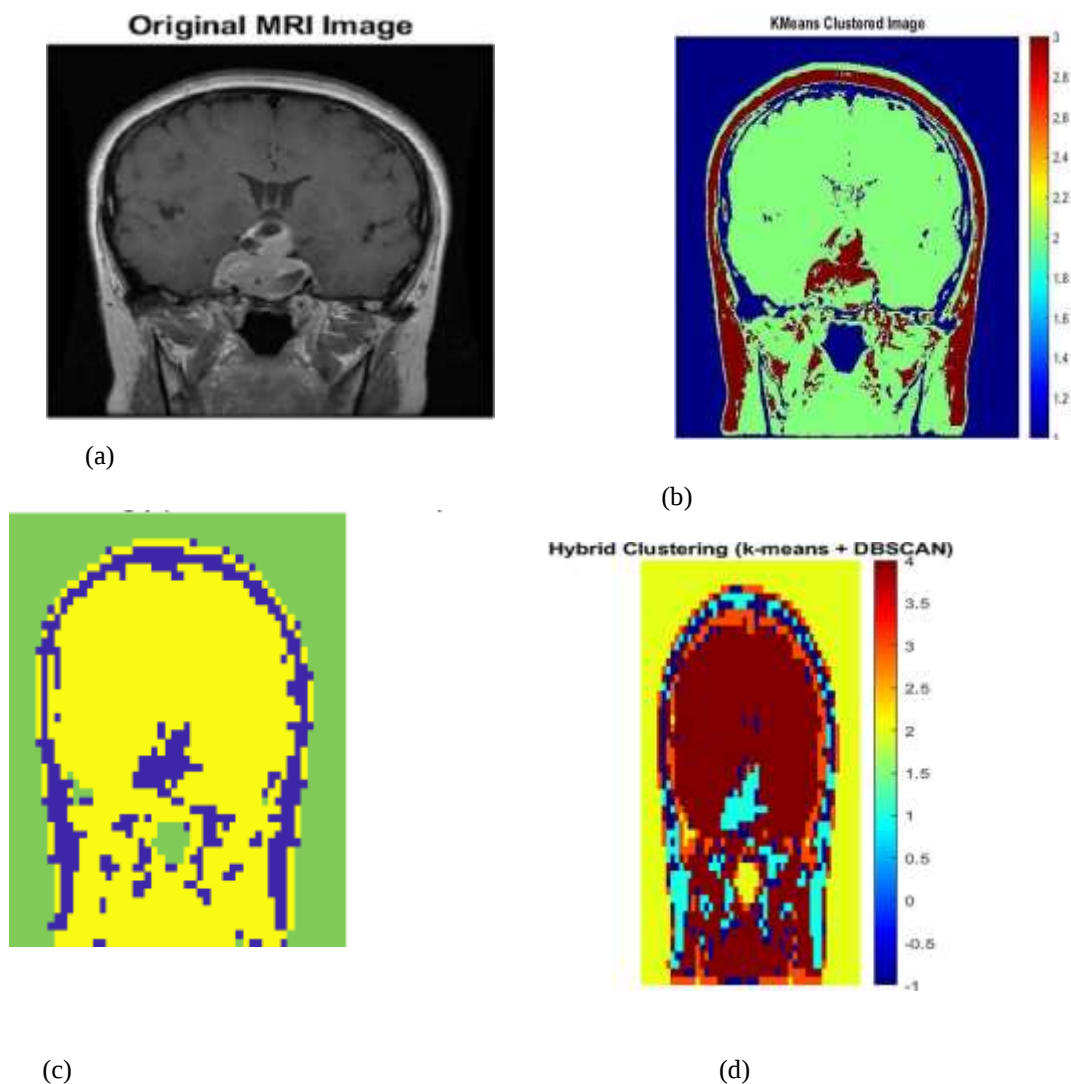


Figure 4: Clustering results on MRI image using different methods: (a) Original MRI Image, (b)K-Means Clustered Image, (c) DBSCAN Clustered Image, (d) Hybrid Clustered Image.



Clustering Algorithm	Silhouette Score	Execution Time
K-Means	0.777614701	0.002718 seconds
DBSCAN	0.611340	0.113359 seconds
Hybrid (K-Means +DBSCAN)	0.800929	0.139168 seconds

Table 1: Average Performance Result of different clustering algorithms on MRI image datasets.

Discussion

The average silhouette score and execution time of K-means, DBSCAN, and a hybrid approach were the main metrics utilized to evaluate clustering methods for MRI image datasets. Table 1 shows the hybrid technique outperformed DBSCAN in this parameter, with K-means and DBSCAN coming in second and third, respectively, according to the data. The hybrid method produced significantly higher silhouette scores than K-means, although taking longer to execute. This indicated that the hybrid method was effective in creating well-defined clusters. Analyzing the clustered images visually showed that, in comparison to K-means, the hybrid method and DBSCAN produced clusters that looked smoother (cluster edges are clearly defined) and extracted more features which is an essential skill for MRI images analysis, as precise diagnostic insights can be obtained by accurately capturing complicated forms as well as controlling noise. Although the hybrid technique required longer execution periods, its improved feature extraction capabilities and higher average silhouette scores illustrate its potential as a useful tool for MRI image clustering. It should be the goal of future research to maximize execution efficiency while maintaining or improving clustering quality, possibly by means of developments in algorithmic approaches or computational optimizations.

CHALLENGES AND RECOMENDATION

One of the significant challenges faced during this research was the absence of ground truth for the image datasets. Ground truth data is essential for validating and evaluating the performance of clustering algorithms. Without it, assessing the accuracy and effectiveness of the clustering results becomes challenging, potentially leading to less reliable conclusions, also the use of (.jpg) images posed another challenge in this research. So, we would recommend future researchers to prioritize using image datasets that come with ground truth annotations, and using NIFTI (.nii) format MRI images for analysis in future studies, (which are images that are specifically designed for storing medical imaging data and retain the full resolution and detail of MRI scans without compression artifacts).

CONCLUSION

In this work, we examined the strengths and weaknesses of the K-means and DBSCAN



clustering methods, as well as their combination, for applications involving MRI image analysis. Our investigation and research reveal that every algorithm has particular benefits and drawbacks. Particularly helpful for complex image data with unusual patterns, DBSCAN excels at managing noise, identifying clusters of any forms, and creating smooth clusters. On the other hand, K-means is a better option for big datasets and real-time applications due to its higher computing efficiency and simplicity. By mitigating the limitations of individual techniques, the hybrid approach provides a more robust and adaptive conclusion. Computational efficiency continues to improve while robust cluster identification is possible with the combination of DBSCAN and K-means. Through the extraction of smoother clusters and more detailed characteristics, this combination technique enhances the accuracy and reliability of MRI image interpretation. According to our research, the hybrid technique combines the advantages of both algorithms to offer a complete solution for challenging image processing tasks. Its versatility makes it a promising method for a range of uses beyond MRI image analysis, such as computer vision, remote sensing, and medical diagnosis. In order to improve clustering performance and scalability, future research should concentrate on enhancing parameter selection, further improving these hybrid methods, and researching integration with deep learning techniques. By addressing these issues, we may create image analysis tools that are more reliable and effective, which will eventually advance the image processing industry.

REFERENCES

- a, a. (n.d.). *Clustering in Machine learning - Image and video recognition*. Retrieved 2024-03-23, from <https://www.linkedin.com/pulse/clustering-machine-learning-image-video-recognition-meghana-bandaru>
- Boulis, C., & Ostendorf, M. (2004). Combining Multiple Clustering Systems. In D. Hutchison et al. (Eds.), *Knowledge Discovery in Databases: PKDD 2004* (Vol. 3202, pp. 63–74). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved 2024-02-25, from http://link.springer.com/10.1007/978-3-540-30116-5_9 (Series Title: Lecture Notes in Computer Science) doi: 10.1007/978-3-540-30116-5_9
- Cluster Analysis*. (2019, May). Retrieved 2024-06-23, from <https://deepai.org/machine-learning-glossary-and-terms/cluster-analysis>
- DBSCAN vs. K-Means: A Guide in Python*. (n.d.). Retrieved 2024-02-27, from <https://www.newhorizons.com/resources/blog/dbscan-vs-kmeans-a-guide-in-python>
- Despotović, I., Goossens, B., & Philips, W. (2015). MRI Segmentation of the Human Brain: Challenges, Methods, and Applications. *Computational and Mathematical Methods in Medicine*, 2015, 1–23. Retrieved 2024-02-26, from <http://www.hindawi.com/journals/cmmm/2015/450341/> doi: 10.1155/2015/450341
- Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996, August). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.. Retrieved 2024-07-03, from <https://www.semanticscholar.org/paper/A-Density-Based-Algorithm-for-Discovering-Clusters-Ester-Kriegel/5c8fe9a0412a078e30eb7e5eeb0068655b673e86>
- Guan, X., Zhao, Y., Nyatega, C. O., & Li, Q. (2023, April). Brain Tumor Segmentation Network with Multi-View Ensemble Discrimination and Kernel-Sharing Dilated Convolution. *Brain Sciences*, 13(4), 650. Retrieved 2024-07-11, from <https://www.mdpi.com/2076-3425/13/4/650> (Number: 4 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/brainsci13040650
- Introduction to Image Segmentation with K-Means clustering*. (n.d.). Retrieved 2024-07-03, from <https://www.kdnuggets.com/introduction-to-image-segmentation-with-k-means-clustering> (Section: 2019 Aug Tutorials, Overviews)
- Jafarzadegan, M., Safi-Esfahani, F., & Beheshti, Z. (2019, December). Combining hierarchical clustering



- approaches using the PCA method. *Expert Systems with Applications*, 137, 1–
10. Retrieved 2024-07-04, from <https://www.sciencedirect.com/science/article/pii/S0957417419304737> doi: 10.1016/j.eswa.2019.06.064
- Jain, R. (2012, May). A Hybrid Clustering Algorithm for Data Mining. *Computer Science & Information Technology*, 2. doi: 10.5121/csit.2012.2239
- Khedairia, S., & Khadir, M. T. (2022, January). A multiple clustering combination approach based on iterative voting process. *Journal of King Saud University - Computer and Information Sciences*, 34(1), 1370–1380. Retrieved 2024-02-25, from <https://linkinghub.elsevier.com/retrieve/pii/S131915781930597X> doi: 10.1016/j.jksuci.2019.09.013
- K-means Clustering Algorithm: Applications, Types, and Demos [Updated] Simplilearn*. (n.d.). Retrieved 2024-02-27, from <https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm>
- Magnetic Resonance Imaging (MRI)*. (n.d.). Retrieved 2024-02-25, from <https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri>
- Nyatega, C. O., Qiang, L., Adamu, M. J., & Kawuwa, H. B. (2022, October). Gray matter, white matter and cerebrospinal fluid abnormalities in Parkinson's disease: A voxel-based morphometry study. *Frontiers in Psychiatry*, 13, 1027907. Retrieved 2024-07-11, from <https://www.frontiersin.org/articles/10.3389/fpsy.2022.1027907/full> doi: 10.3389/fpsy.2022.1027907
- Nyatega, C. O., Qiang, L., Adamu, M. J., Younis, A., & Kawuwa, H. B. (2021, December). Altered Dynamic Functional Connectivity of Cuneus in Schizophrenia Patients: A Resting-State fMRI Study. *Applied Sciences*, 11(23), 11392. Retrieved 2024-07-11, from <https://www.mdpi.com/2076-3417/11/23/11392> doi: 10.3390/app112311392
- Pedro, J. (2022, May). *How to ensemble Clustering Algorithms*. Retrieved 2024-02-28, from <https://towardsdatascience.com/how-to-ensemble-clustering-algorithms-bf78d7602265>
- Sharma, E. (2020, June). *K-Means vs. DBSCAN Clustering*. Retrieved 2024-07-03, from <https://towardsdatascience.com/k-means-vs-dbscan-clustering-49f8e627de27>
- Strehl, A., & Ghosh, J. (n.d.). Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions.
- Velunachiyar, S., & Sivakumar, K. (2023, June). Some Clustering Methods, Algorithms and their Applications. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(6s), 401–410. Retrieved 2024-02-26, from <https://ijritcc.org/index.php/ijritcc/article/view/6946> doi: 10.17762/ijritcc.v11i6s.6946
- Yan, J., & Liu, W. (2022, May). An Ensemble Clustering Approach (Consensus Clustering) for High-Dimensional Data. *Security and Communication Networks*, 2022, 1–9. Retrieved 2024-02-26, from <https://www.hindawi.com/journals/scn/2022/5629710/> doi: 10.1155/2022/5629710