Machine Learning-Based Gene Clustering on Brain Cancer Using K-Means and Hierarchical Clustering Methods

Fatih Yilmaz*, Samed Jukic*
*International Burch University
fatih.yilmaz@stu.ibu.edu.ba
samed.jukic@ibu.edu.ba

Original research

Abstract: *K-means and hierarchical clustering algorithms are employed to cluster genes according to the gene expression to determine the harming level of the genes in brain cancer. The gene expression data with a control group from The Cancer Genome Atlas database were used. The optimal cluster number for each clustering technique was obtained using the elbow method and dendrogram for K-means and hierarchical clustering methods respectively. We identified the ideal number of clusters as three and further classified them into seven groups. We observed that the second cluster contains over half the genes in healthy people and the cluster distribution of a healthy patient and a patient who died six months after being diagnosed with brain cancer is similar. Further analysis indicated that of all the time spent by patients after being diagnosed with brain cancer, group 0 has the highest percentage in one month after the diagnosis, while group -2 has the lowest percentage. Most genes shift their clusters when K-means and hierarchical clustering techniques we compared with the genes from the control and disease groups. The result of the measure of dissimilarity between the genes expression patterns indicates that the K-means technique outperforms the hierarchical technique with a higher rate of change in the cluster.*

Keywords***: Brain cancer, gene clustering, hierarchical clustering, K-means clustering, machine learning.***

1. Introduction

The brain is the most important part of the body because it controls the central nervous system of the human body. It takes charge of several actions of the body, thereby playing a key role in the human nervous system. One of the leading causes of death and a major obstacle in increasing life expectancy in the world is cancer [1]. A World Health Organization (WHO) report in 2019 indicated that cancer is the foremost or the second prominent cause of death before attaining the age of seventy in 112 countries of the world [2]. Brain cancer is one of several types of cancer diseases, and it develops in the brain [3]. Some warning signs of brain cancer amongst others include frequent headaches, speech changes, coordination problems, and memory loss. This type of cancer stays in the brain [4]. Categories of brain cancer are based on the development stage, origin, growth rate, and nature. The cancer of the brain can be of two types, either malignant or benign [4]. The malignant brain cancer cells attack nearby cells present in the spinal cord or brain; they have fast development rates. Benign brain cancer cells hardly attack the nearby healthy cells, they exhibit slow development rates and have distinctive borders. Brain cancer can be diagnosed either invasively or non-invasively. The invasive technique includes doing a small opening to collect cancer samples for necessary clinical tests, where the samples are subjected to microscopic examination to ascertain the malignancy. The non-invasive technique includes carrying out a physical examination of the brain and the entire body using imaging systems such as magnetic resonance imaging and computed tomography, which are quicker and harmless than the invasive approach. The imaging methods enable the radiologists to identify brain disorders, observe the pattern of development and assist in surgical preparation [5]. The introduction of dominant computing machines has led to the decrease in the cost of diagnostic hardware through the development and deployment of computer-aided tools for brain cancer diagnosis. These tools are projected to enhance radiologists' ability to accurately and consistently deliver quality diagnostic results. In this study, two machine learning models; hierarchical clustering and K means were developed to cluster the gene present in brain cancer. Clustering algorithms are the main computational tools employed in this study. Clustering analysis includes the process of data grouping into two or more clusters such that data points in the same cluster are like those in different clusters due to information retrieved through the data points [6]. Carrying out clustering analysis on groups of cancer samples having similar patterns can lead to the discovery of new cancer subtypes. Clustering analysis was first employed in the study, "Molecular Classification of Cancer" [7].

A. Clustering Techniques

Clustering techniques were used after the preceding procedures were completed properly. The following sections go over two different clustering techniques:

B. K-means Clustering

K-Means clustering aims to segregate data into groups, and usually, the grouping is occasionally characterized by the variable 'k'. The algorithm makes an effort to assign each information point to the variable 'k' groups available concerning the feature similarity. This method of clustering data is usually appropriate for use because it is relatively simple to implement the variables and at the same time generalize clusters of distinct shapes and sizes like the elliptical. It can also be used to scale large data sets, which saves time, and reduces the tediousness of the grouping. Additionally, K-Means clustering adapts quickly to new examples making it easy to understand and interpret hence it is useful because of its flexibility. Thus, in summary, K-means clustering is a technique used in objects in a procedure that minimizes their variation amongst them. Among the various types of existing K-mean algorithms, the one in

practice highlights the variation present in a group as the summation of the squared distances of Euclidean existing between each element of the group and centroid ad is given as

$$W(C_k) = \sum_{x_i \epsilon} (x_i - \mu k)^2 \quad (1)$$

In equation (1), x_i is the data that belongs to the C_k cluster, and μk is the mean value of the data given to the C_k cluster. [8]. Thus, the sum of all the K clusters divided by the summation of the Euclidean distances is then squared amongst the data and to the corresponding center.

C. Hierarchical Clustering

Ultimately, hierarchical clustering is another method of grouping data that seeks to set up a hierarchy of clusters that usually are comprised of two types; the Agglomerative type also known as the bottom-up approach ensures each observation begins in their clusters and a pair of clusters are brought together moving up the hierarchy. Divisive also known as the top-down approach ensures all the observations begin in one cluster. The splits are carried out recurrently as they move down the hierarchy. Notably, the results of this clustering method are always presented in the form of a dendrogram. Hierarchical clustering is beneficial because it is easy to implement and usually assures the best results in most of the areas it is applied to. Furthermore, there is always no specific information about the number of clusters required, making it suitable for every application [9].

2. Literature Review

The incessant process of unwanted death of cells apart from the production of new ones is controlled by the genes. The development of cancerous cells originates from uncontrolled cell growth. Medical imaging methods have helped health professionals and researchers to have a deeper view of the inner human body and carry out the analysis of this part without undergoing incisions. To accord proper cancer treatment, diagnosis, grade assessment, treatment response assessment, surgery planning, and patient prognosis are the key steps to follow. There are two brain imaging approaches, and they are functional and structural imaging [10]. The functional approach identifies the metabolic changes, cuts on an improved scale, and visualizes the activities of the brain. On the other hand, the structural approach includes several measures associated with brain cancer location, structure, injuries amongst others. MRI and CT are mostly used for brain cancer analysis which can capture various sections of the body without any surgery [11]. To further improve brain image analysis, several machine learning models have been used in describing brain tumors [12]. In the application of machine learning for brain image characterization, two important stages are involved: feature extraction and classification. The feature extraction stage involves a set of mathematical models that are built around some image characteristics such as contrast, texture, and brightness. To improve the perceived power of the model, several features accrued from various extraction models are joined together [13]. Brain images are classified and segmented using models such as Artificial Neural Network, K-Nearest Neighbors, Support Vector Machines, metaheuristic algorithms, region growing algorithms, and morphology amongst others.

In biomedicine, cluster analysis is a major data mining approach that is employed in data analysis processes. The hierarchical clustering technique is a classical clustering approach that has been broadly used in the field of biomedical. According to this study at BMC Bioinformatics, the major reason for using hierarchical clustering is its simplicity. It requires just one parameter, the number of clusters, and the availability of implementations as part of the software [14]. Another classical clustering algorithm is K-means, which is a method that requires cluster numbers to be given as input by the user. Generally, finding a suitable value for the cluster

numbers is a demanding task [15]. K-means has been identified as one of the best clustering algorithms used for analyzing cancer gene expression data [16], even with its non-deterministic nature issue, which is one of the main drawbacks of the K-means technique. Also, the K-means algorithm is relatively simple to implement [17].

3. Methodology

Our study can be divided into the following parts:

- Data collection
- Preprocess and prepare the dataset
- Employing clustering techniques

A flowchart of our overall analysis has been shown in Figure 1.

A. Data Collection

The microarray gene expression values used in this research were derived from the TCGA gene expression database. Six death groups, alive and one control group included in the dataset were studied. The groups studied are one month, three months, six months, one year, two years, three years later and alive. They include 20, 34, 76, 99, 144, 73, and 147 patients respectively. All six groups comprised of the dead, alive, and the control group had the same number of 17814 genes for the analysis. All groups had ID, death time an average of gene expressions, and gene ID.

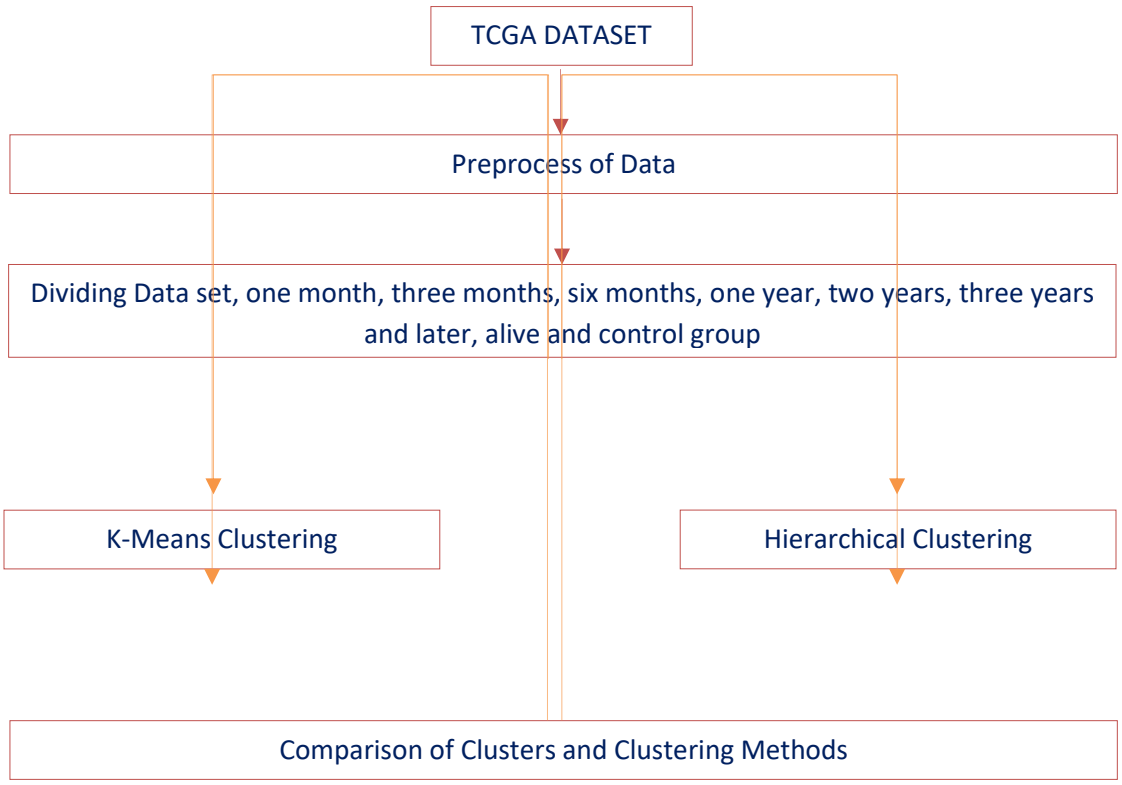


FIGURE 1. Flow-chart of our overall study.

B. Preprocess and Preparing Dataset

In the initial stage, patients whose data was missing in the clinical dataset were removed. Dead people's and alive patients' information were taken from the data set and dead people were classified into six different groups which were one month, three months, six months, one year, two years, and three years later depending on the duration of death time. Finally, other groups comprising of deceased, alive, and control groups were obtained.

4. Results and Discussion

A. Optimal Number of Clusters Calculation

To evaluate the optimal number of clusters for the clustering techniques, the Elbow method was used for K-means techniques while dendrogram was used for hierarchical techniques.

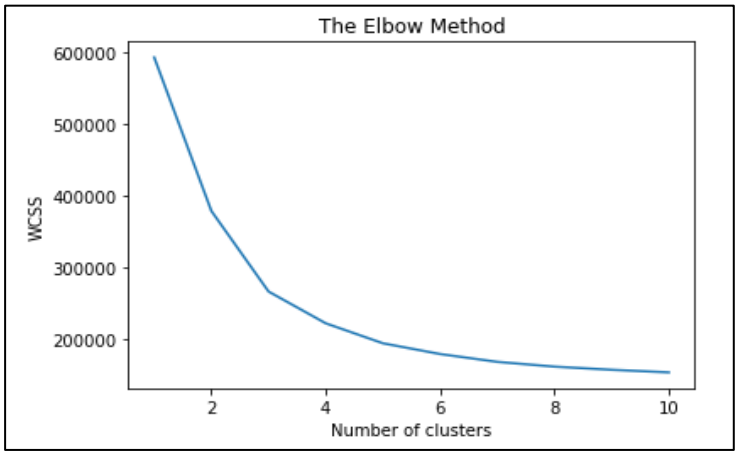
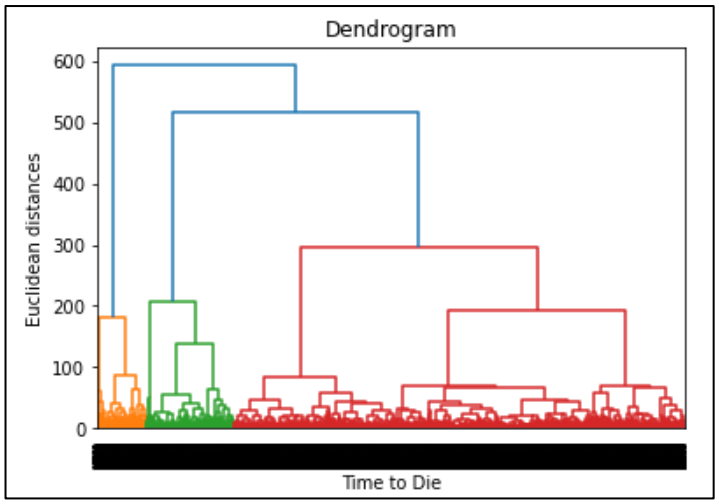


FIGURE 2. Elbow Method Curve for K-Means technique (for one month dataset).



3 clusters

FIGURE 3. Dendrogram for hierarchical clustering technique (for one month dataset).

B. *Dataset Evaluation*

The given graph in Fig. 4 shows a two-dimensional representation of the dataset which clustered the whole dataset that has 17814 genes separated into three groups using K-means clustering methods. These groups are 0th, 1st, and 2nd clusters. 69.32% of the genes in the control group belong to the 2nd cluster, 12.61% to the 1st cluster, and 18.07% to the 0th cluster (Table 1).

TABLE 1. Distribution Genes in the Clusters using K-Means Clustering Method.

Cluster	One Month	Three Months	Six Months	One Year	Two Years	Three Years	Alive	Control
0-Cluster	3578	12057	3420	11921	2361	11789	12088	3219
1-Cluster	11941	3445	2338	2307	11898	2240	3379	2246
2-Cluster	2295	2312	12056	3586	3555	3785	2347	12349

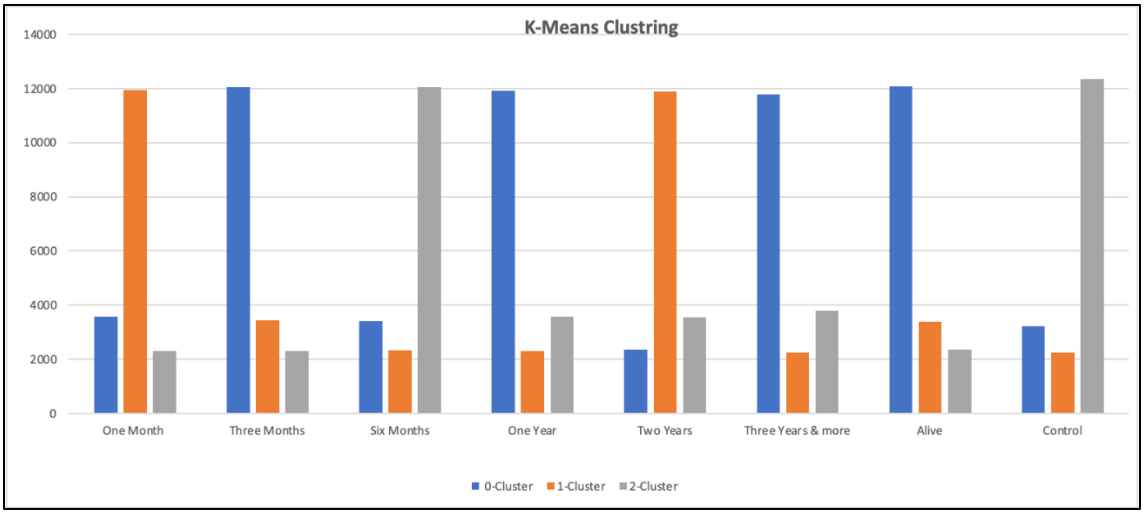


FIGURE 4. Clustering Dataset using K-Means Clustering Method.

From this graph, it can be observed that the second cluster contains more than half of the genes in healthy people. It has been determined that the cluster distribution of a healthy patient and a patient who died six months after being diagnosed with brain cancer is similar. Approximately 67% of the genes of healthy people and six months groups at the second cluster.

C. *Model Evaluation*

The developed gene expression clustering models were evaluated based on certain performance metrics such as classification based on min-max and average value. All these metrics were evaluated to ascertain the performance of the gene clustering model.

D. Min-Max Gene Expression Outcome

The clustering of gene expression data of brain cancer patients was evaluated in two-phase. The first phase of the clustering process was done by classifying each gene expression of each patient into three main outcomes: low, normal, and high based on their equivalent biosample ID and the aggregated minimum and maximum values of each type of gene as presented in Tables 2.

TABLE 2. Classification based on Min-Max Value.

Input	Rule	Outcome
Biosample repository ID	Value of biosample repository ID>Max	Low
Biosample repository ID	Value of biosample repository ID >Min, Value of biosample repository ID <Max	Normal
Biosample repository ID	Value of biosample repository ID >Max	High

E. Gene Expression Classes

The dataset was further divided into seven different groups based on the time spent by the patients after diagnosis, this was achieved by taking their arithmetic means. The average value of each gene was calculated by taking the average of the expression data, the outcome was classified into different groups. Any average value that is 0, belongs to class 0. The average value of 1, belongs to class 1. The average value of 2, belongs to class 2. The average value of -1, belongs to class -1. Lastly, an average value of -2, and belongs to class 2. In a case of greater than 2 average value, that is an extremely high class, and an average value greater than -2 is extremely low.

F. Gene Expression Classification

One of the objectives of this work is to classify the gene expression of brain cancer patients into three main (0, 1, 2) clusters of seven groups either 0, 1, -1, 2, -2, extremely low, extremely high. Fi 5 presents the percentage graphical representation of all the time spent by patients after being diagnosed with brain cancer in each class. Group 0 has the highest percentage in One month after the diagnosis dataset. Group -2 has the lowest percentage.

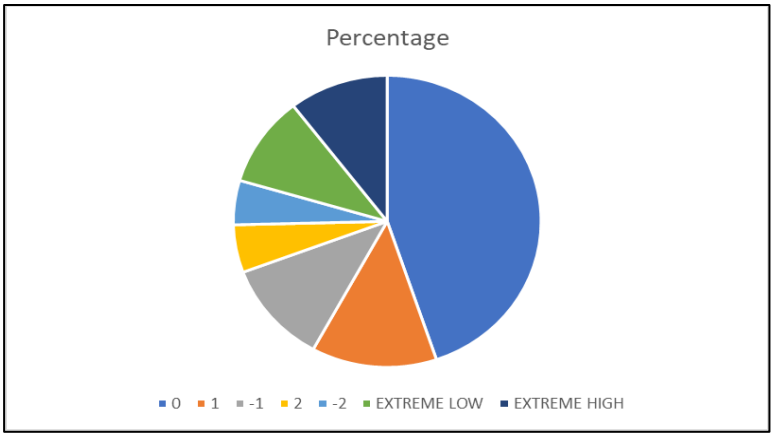


FIGURE 5. Gene Expression Classification for Patient Diagnosed with Brain Cancer One Month after Diagnosis.

G. Cluster Validity

This involves evaluating the clustering analysis results in a quantitative method. Clustering validity metrics are usually employed to calculate the optimal cluster numbers and it is usually dependent on clustering techniques employed. The idea behind cluster validity is to discover changing and non-changing clusters also known as compact and well-separated clusters. Compactness is employed to measure data variation present in a particular cluster, while the separation denotes segregation of clusters from one another. To achieve cluster validity, validity measures use sample mean of each subset, while others use all the points present in each subset in their computation.

H. K-Means Cluster Changing Rate

K-means clustering technique computes the centroids and iterates till optimal centroid is evaluated. The rate of change of cluster in the k-means technique is faster due to the quick identification of k number centroids that allocates every data point to the nearest cluster. The cluster comparison to control for clusters 0, 1, and 2 is 78.17, 8.77, and 13.06 respectively. This was evaluated by taking the percentage of a total number of genes in control for each cluster: 13925, 1563, 2326 respectively divided by a total number of genes.
Table 3 illustrates the cluster change for patient death duration after one month, each of the individual clusters has relatively the same cluster change pattern. For one month dataset using k-means clustering techniques, the cluster change rate is faster. The total number of genes present in the one-month dataset for cluster 0 is 3578, cluster 1 is 11941, and cluster 2 is 2295.

TABLE 3. One Month Death Duration Cluster Changing Rate for K-Means Technique.

Death Duration	**Method of Clustering**	**Clusters**	**Total Genes**	**Cluster Changing**
One Month	K-Means	0	3578	Yes
		0		No
		1	11941	Yes
		1		No
		2	2295	Yes
		2		No

I. Hierarchical Cluster Changing Rate

The cluster comparison to control for clusters 0, 1, and 2 are 78.17, 8.77, and 13.06 respectively. This was evaluated by taking the percentage of a total number of genes in control for each cluster: 13925, 1563, 2326. Presented in Table 4 is the cluster change for patient death duration after one month, each of the individual clusters has relatively the same cluster change pattern. For one month dataset using hierarchical clustering techniques, the cluster change rate is less fast compared with the K-means clustering technique. The total number of genes present in the one-month dataset for cluster 0 is 13679, cluster 1 is 1490, and cluster 2 is 2645.

The bar graph in Fig. 5 shows which group changed cluster depending on the control group. Compared to the genes in the control group and one month, three months, one year, two years, three years later, and alive groups have changed clusters from 72% to 79%. In the six months, just 47% of the genes had changed the cluster. The alive group has the highest rate at 78.89% in gene cluster replacement rates. The six-month group has the lowest rate, at 47.32%.

TABLE 4. One Month Death Duration Cluster Changing Rate for Hierarchical Technique.

Death Duration	Method of Clustering	Clusters	Total Genes	Cluster Changing
One Month	Hierarchical	0	13679	Yes
		0		No
		1	1490	Yes
		1		No
		2	2645	Yes
		2		No

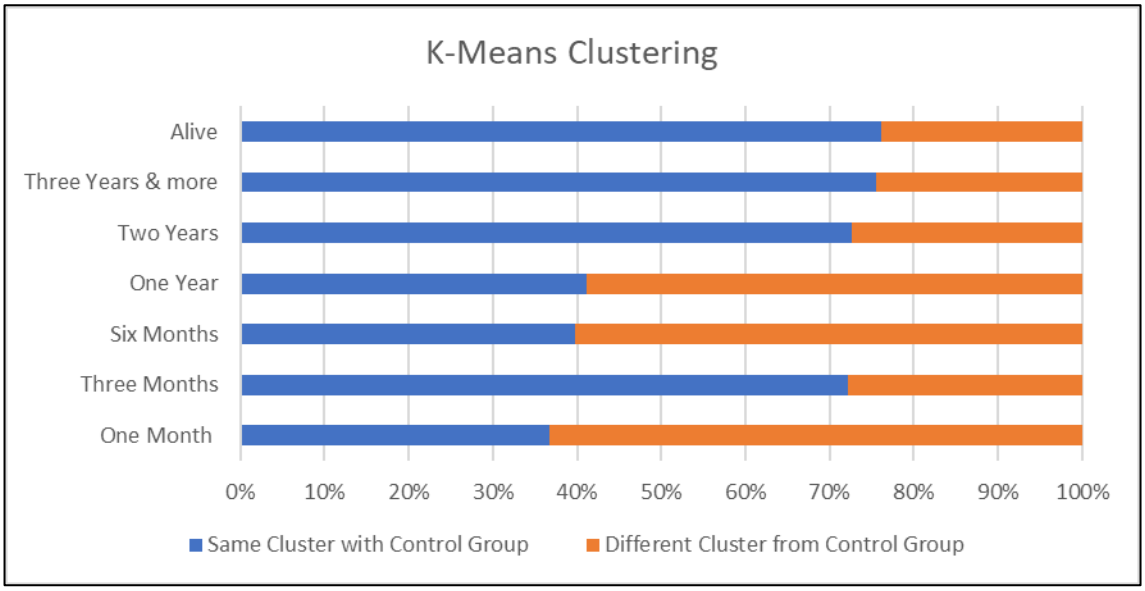


FIGURE 5. Shifting cluster between the control group and diseased groups.

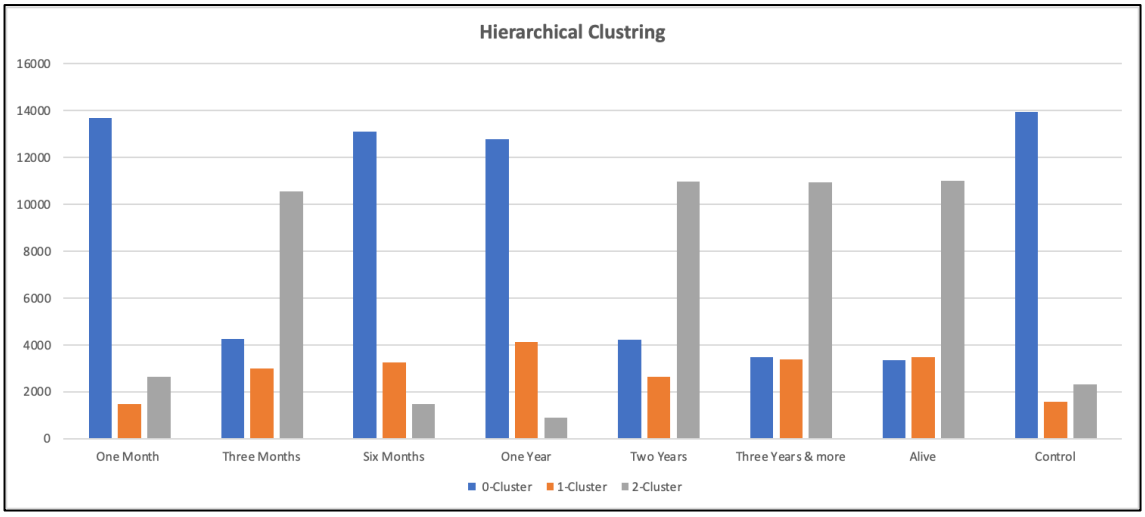


FIGURE 6. Clustering dataset using by Hierarchical Clustering Method.

The graph shows the distribution of genes using the Hierarchical Clustering method (Figure 6). The distribution of people in the control group using the Hierarchical Clustering method is in the table. 13925 of the genes are in the 0th cluster, 1563 are in the 1st cluster, and 2326 are in the 2nd cluster. In a healthy individual, the 0th cluster is seen as dominant. When the groups are observed, one month, six months, one year, and the control group have a similar distribution of genes, approximately 70% of genes in the 0th cluster. When the three months, two years, three years, and alive groups are examined, between 50% and 60% of the genes are in the 2nd cluster.

Table 5 shows the results of the hierarchical clustering method, including the number of gene clusters that have changed and how many of these are in the same cluster as their control group. According to the statistics in the table, the alive group had the greatest rate of people who changed clusters, at 76.19%. The one-month group had the smallest change of 36.80%.

TABLE 5. Distribution genes in the Clusters using Hierarchical Clustering Method.

Cluster	One Month	Three Months	Six Months	One Year	Two Years	Three Years & more	Alive	Control
0-Cluster	13679	4260	13105	12777	4220	3492	3346	13925
1-Cluster	1490	3001	3241	4121	2633	3376	3476	1563
2-Cluster	2645	10553	1468	916	10961	10946	10992	2326

6. Conclusion

Using clinical data and gene expression data from its database, seven distinct groups, and control groups were formed in this study. In the formed groups, three different clusters were obtained by using K means and Hierarchal clustering methods. In these two approaches, three clusters were also used to create gene expression in the control group. Most of the genes in the control group were in the second cluster. On the other hand, according to the data obtained using the K-means clustering approach, most of the genes in patients with the disease differed in the cluster distribution. In the result of the Hierarchical Clustering Method, most of the genes in the control group were in the 0th cluster and three groups were found to have a similar distribution to clusters in the control group. Except for these three groups, the gene distribution of the other groups differed. When K-means Clustering Method was compared with the Hierarchical Clustering Method, it could be found that the cluster change rate according to the control group was higher in the K-means Clustering Method. In the future, work can be done using different machine learning clustering methods. By using Clustering methods and extreme gene expressions, it can be revealed which genes are effective in brain cancer.

7. References

[1] World Health Organization (WHO). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. WHO; 2020. [Accessed 16 June 2021]. who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

[2] Amyot, F., Arciniegas, D. B., Brazaitis, M. P., Curley, K. C., Diaz-Arrastia, R., Gandjbakhche, A., Herscovitch, P., Hinds, S. R., Manley, G. T., Pacifico, A., Razumovsky, A., Riley, J., Salzer, W., Shih, R., Smirniotopoulos, J. G., & Stocker, D. (2015). A Review of the Effectiveness of

Neuroimaging Modalities for the Detection of Traumatic Brain Injury. *Journal of Neurotrauma*, *32*(22), 1693–1721. https://doi.org/10.1089/neu.2013.3306

[3] Bray, F., Laversanne, M., Weiderpass, E., & Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, *127*(16), 3029–3030. https://doi.org/10.1002/cncr.33587

[4] de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: A comparative study. *BMC Bioinformatics*, *9*(1), 497. https://doi.org/10.1186/1471-2105-9-497

[5] Dirks, P. B. (2008). Brain Tumor Stem Cells: Bringing Order to the Chaos of Brain Cancer. *Journal of Clinical Oncology*, *26*(17), 2916–2924. https://doi.org/10.1200/JCO.2008.17.6792

[6] Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine Learning for Medical Imaging. *RadioGraphics*, *37*(2), 505–515. https://doi.org/10.1148/rg.2017160130

[7] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, *286*(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

[8] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[9] Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Edition 1). STHDA.

[10] Mahaley, M. S., Mettlin, C., Natarajan, N., Laws, E. R., & Peace, B. B. (1989). National survey of patterns of care for brain-tumor patients. *Journal of Neurosurgery*, *71*(6), 826–836. https://doi.org/10.3171/jns.1989.71.6.0826

[11]Morris, Z., Whiteley, W. N., Longstreth, W. T., Weber, F., Lee, Y.-C., Tsushima, Y., Alphs, H., Ladd, S. C., Warlow, C., Wardlaw, J. M., & Al-Shahi Salman, R. (2009). Incidental findings on brain magnetic resonance imaging: Systematic review and meta-analysis. *BMJ*, *339*(aug17 1), b3016–b3016. https://doi.org/10.1136/bmj.b3016

[12] Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: An overview, II. *WIREs Data Mining and Knowledge Discovery*, *7*(6). https://doi.org/10.1002/widm.1219

[13] Nidheesh, N., Abdul Nazeer, K. A., & Ameer, P. M. (2018). *A Hierarchical Clustering Algorithm Based on Silhouette Index for Cancer Subtype Discovery from Omics Data* [Preprint]. Bioinformatics. https://doi.org/10.1101/309716

[14] Tan, P.-N., Steinbach, M., & Kumar, V. (2006a). *Introduction to data mining* (1st ed). Pearson Addison Wesley.

[15] Tan, P.-N., Steinbach, M., & Kumar, V. (2006b). *Introduction to data mining* (1st ed). Pearson Addison Wesley.

[16] Tan, P.-N., Steinbach, M., & Kumar, V. (2010a). *Introduction to data mining* (Pearson internat. ed., [Nachdr.]). Pearson Addison-Wesley.

[17] Tan, P.-N., Steinbach, M., & Kumar, V. (2010b). *Introduction to data mining* (Pearson internat. ed., [Nachdr.]). Pearson Addison-Wesley.

[18] Tandel, G. S., Biswas, M., Kakde, O. G., Tiwari, A., Suri, H. S., Turk, M., Laird, J., Asare, C., Ankrah, A. A., Khanna, N. N., Madhusudhan, B. K., Saba, L., & Suri, J. S. (2019). A Review on a Deep Learning Perspective in Brain Cancer Classification. *Cancers*, *11*(1), 111. https://doi.org/10.3390/cancers11010111

[19] VASANTHA, M., SUBBIAH BHARATHI, & SUBBIAH BHARATHI. (2010). Medical Image Feature, Extraction, Selection And Classification. *International Journal of Engineering Science and Technology*, *2*(6), 2071–2076.