Prediction of Cancer in DNA Sequences Using Unsupervised Learning Methods

Seyma Dogru [1*], **Volkan Altuntas**[2]

[1*,2] *Department of Computer Engineering, Bursa Technical University, 16310 Bursa, Turkey*

Abstract

Today, with the development of technology, the decision-making capabilities of machines have also increased. With their high analytical skills, computers can easily catch points and relationships that may escape the human eye.

Thanks to these capabilities, machines are also widely used in the field of health. For example, many machine-learning techniques developed on cancer prediction have been successfully applied. Early detection of cancer is crucial to survival. In the early diagnosis of cancer, the rates of drug treatment, chemotherapy, or radiotherapy that the person will be exposed to are significantly reduced and the patient gets through this process with the least amount of wear and tear. Gene Expression Cancer RNA-Seq Dataset was used in this study. This data set includes gene expression values of 5 cancer types (BRCA, KIRC, LUAD, LUSC, UCEC). DNA sequences in the dataset were analyzed using k-means and hierarchical clustering algorithms, which are unsupervised machine learning methods. The aim of the study is to develop a usable machine-learning model for the early detection of cancer at the gene level. Adjusted Rand Index (ARI), Silhouette Score, and Accuracy Metrics were used to evaluate the analysis results. The rand index calculates the similarity between clusters by counting the binaries assigned to clusters. The adjusted Rand Index is a randomly adjusted version of the Rand Index. The silhouette score indicates how well a data point fits within its own set among separated datasets. The accuracy metric is obtained as a percentage of correctly clustered data points divided by all predictions. Different connection methods are used in the hierarchical clustering algorithm. These are 'complete', 'ward', 'average', and 'single'. As a result of the study, the accuracy in the k-means algorithm was 0.990, the Adjusted Rand Index was 0.79, and the Silhouette Score was 0.14. Looking at the hierarchical clustering, ward performed the best of the four linkage methods, with an ARI score of 0.76 and a silhouette score of 0.13. As a result of the study, the accuracy of the hierarchical clustering algorithm was 0.999.

Keywords: k-Means, Machine Learning, Unsupervised Learning, Cancer, DNA, Hierarchical Clustering.

Cite this paper as:
Dogru, S. and Altuntas, V. (2023). *Prediction of Cancer in DNA Sequences Using Unsupervised Learning Methods*. Journal of Innovative Science and Engineering.7(1):40-47

*Corresponding author: Seyma Dogru
E-mail: seymadogru37@gmail.com

Received Date:05/07/2022
Accepted Date:17/10/2022

1. Introduction

Cancer is a disease in which certain cells in the body grow uncontrollably and spread to other parts of the body. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. In a healthy human body, cells grow and then multiply to form new cells. They also die when they take damage or get old. Instead, the body produces new cells. In some bodies, this order can be disrupted, and damaged cells grow and multiply abnormally when they shouldn't. As a result, this disease, which we call cancer, occurs. As in all diseases, early diagnosis of cancer is very important to save the life of the patient. Many studies have been conducted over the years to facilitate diagnosis. Cancer can be diagnosed at the gene level and with imaging techniques. Cancers are heterogeneous due to their habitat. Subtypes of cancer types can also be found. For example, there are 5 subtypes of cancer for breast cancer. It has been named Luminal A, Luminal B, HER2 overexpressed, Basal-like, and Claudin-low [1]. In this study, the diagnosis of cancer subtypes by DNA sequencing was investigated. Each subtype has a unique treatment plan, and in this respect, a successful diagnosis of the subtype is important for the patient to receive the right treatment. In previous studies, supervised learning algorithms of machine learning method (Naive Bayes, Decision Tree, Random Forest, Support Decision Machine, K Nearest Neighbor) were used. There are also studies using the Hierarchical Clustering Algorithm [2, 3]. The hierarchical clustering algorithm was chosen by the researchers because of the inherent complexity of most other algorithms and the difficulty of determining the parameters that need to be determined when applying. In this study, k-Means and Hierarchical Clustering algorithms, which are unsupervised machine learning algorithms, were successfully applied and it was observed how they performed on unlabeled data. In the k-Means algorithm, the regions where the data points are concentrated are taken as the center. Data points around these centers are included in the clusters using the Euclidean distance. On the other hand, hierarchical clustering was made, and clustering was done with four linking methods. Single, ward, complete and average were used, and their results were compared. It has been shown that both algorithms give successful results in cancer detection at the gene level. 99.0% accuracy in the k-Means algorithm and 99.9% accuracy in the hierarchical clustering algorithm was obtained.

2. Literature Review

In the study conducted by Fahad Hussain and his colleagues in 2019, cancer prediction was made on the “Gene expression cancer RNA-Seq Data Set” dataset using Naive Bayes, Decision Tree, Random Forest, Support Decision Machine, K Nearest Neighbour algorithms [4]. Elaheh Moradi and colleagues conducted a study titled "Machine learning framework for early MRI-based Alzheimer's transformation prediction in MCI subjects." They used MRI images to identify Alzheimer's disease 1-3 years before clinical diagnosis. For this purpose, semi-supervised and supervised machine learning algorithms were used. [5]. In the study of Konstantina Kourou and her colleagues in 2015, various machine learning and data mining methods were applied to solve the diagnosis problem of cancer patients. The authors conducted a qualitative survey of research articles over the past five years and identified several studies such as cancer susceptibility, cancer recurrence, and cancer survival [6]. In the study conducted by Gunasekaran Manogaran and his colleagues in 2018, Bayes Markov Model and Gaussian Clustering were used to model DNA copy number variation across the genome and identification of cancer [7]. In the study conducted by Zeid Khitan and his colleagues in 2017, algorithms such as SVM, Decision Tree, Neural Network, Random Forest, Linear Model were used to predict the outcome of dialysis, death, or creatinine elevation in kidney patients due to diet [8]. In the study conducted by Manish

Motwani and his colleagues in 2017, they used machine learning algorithms to analyze the mortality within 5 years in patients who underwent angiography [9]. A lung cancer prediction model was developed by Timor Kadir and Fergus Gleeson in 2018 using machine learning and imaging techniques [10]. In 2020, Milon Islam and his colleagues created a prediction model on breast cancer using machine learning techniques. They achieved an accuracy of over 98% in ANN, while they achieved an accuracy of 97.14% in SVM [11]. In the study by Yixuan Li and Zixuan Chen in 2018, different machine learning methods on breast cancer (Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), and Logistic Regression (LR)) have been tested and their performances compared. Accuracy was 0.961, 0.951, 0.961, 0.937, and 0.956 for DT, SVM, RF, RF, LR, and NN, respectively [12]. In the study "Using machine learning to predict ovarian cancer" by Mingyang Lu et al., ovarian cancer prediction was made using algorithms such as Decision tree and Logistic regression [13].

3. Method

In this article, the k-Means algorithm, a machine learning algorithm, and a hierarchical clustering algorithm were used to analyze the DNA sequences of cancer patients. The K-Means algorithm is one of the most popular unsupervised machine learning algorithms. Unsupervised machine learning algorithms work based on clustering.

3.1. Data Set

DNA data of cancer patients were obtained through the data set named “Gene expression cancer RNA-Seq Data Set” [14]. This data set contains 20531 features and 801 rows of data. These classes represent the gene sequence for patients with five types of tumors. Dataset was used in this study. This data set includes gene expression values of 5 cancer types (BRCA, KIRC, LUAD, LUSC, UCEC). Cancer, designated BRCA, is the most aggressive type of breast cancer seen in women. KIRC is a type of kidney cancer known for its high mortality rate worldwide. An average of 75 percent of kidney cancers are diagnosed with this type of cancer. LUAD cancer, which accounts for 40 percent of lung cancer diagnoses, is known to spread more slowly than other types of lung cancer. It is also seen in non-smokers. LUSC, the second most common form of lung cancer, is the most common type of cancer among smokers. Their location is usually in the middle of the lung. UCEC is a type of cancer encountered before birth.

The mortality rate is very high due to the difficulty of early diagnosis. It is known as the most common cancer in women [15].

3.2. K-Means

The clustering technique is only one of the most preferred data analysis methods to obtain information about the classes of data. It can be defined as the task of identifying subgroups so that the data in the same cluster are very similar, while the data in different clusters are very different. Based on a similarity measure, such as Euclidean distance or correlation-based distance, an attempt is made to find homogeneous subgroups within the data so that the data points in each cluster are as similar as possible. The decision of which similarity measurement method to use is application specific [16]. Typically, unsupervised algorithms arrive at a conclusion from datasets using only input vectors, without reference to known or labeled results [17]. It aims to put data showing similar characteristics in the same cluster. There is no clustered data. Data that does not belong to any cluster is included in the Outlier cluster.

K-means, a clustering algorithm, determines subsets of unlabeled data points. The k parameter requested from the user specifies how many subsets it should be divided into. The algorithm decides which dataset will be in which cluster, taking into account the similarity of the data points, without the need for prior training. When the algorithm starts working, it assigns centroids as many as the number of clusters we have determined, so it works on a centroid basis.

The algorithm in question tries to minimize the sum of the differences between the data and the clusters in which they are located. It takes unlabeled points as input data and splits them into k clusters. Iterations continue until the best separation is found.

There are some metrics to measure the success of the study. These are adjusted rand score (ARI), silhouette score, and accuracy. In the ARI score, "0" indicates random labeling, and "1" indicates a perfect replica of the truth. If the silhouette is in the score value, the values are interpreted as -1 for very bad clusters and +1 for very dense clusters.

The Accuracy formula for k-Means is calculated using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) metrics as follows.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \tag{1}$$

3.3. Hierarchical Clustering

The hierarchical clustering method was also used in the study. Hierarchical Clustering was developed to eliminate a disadvantage in the k-Means algorithm. In the k-Means algorithm, the number of clusters must be specified beforehand. For this reason, hierarchical clustering has emerged. The basic logic of the hierarchical clustering algorithm is based on the combination of similar features or vice versa. There are two basic approach logics; agglomerative and divisive. In Agglomerative logic, first of all, all data is separate from each other, so all data is considered as a set on its own. Then, by looking at the characteristics of each, data with similar attributes are started to be thrown into the same cluster. In the Divisive method, the opposite of this process is performed. First of all, all the data is looked at as if it were a cluster and it is started to be divided into subsets according to their distance or similarity. As a result of this process, each data becomes a cluster on its own.

Different linkage methods are used while performing these operations. Linkage methods are Single Linkage, Complete Linkage, Average Linkage, Ward's Linkage, Median Linkage, and Centroid Linkage. Among these, Single Linkage, Complete Linkage, and Average Linkage methods are link-based techniques. Ward's Linkage method is variance based. Median Linkage and Centroid Linkage are based on centralization. Single Linkage proceeds by merging the two closest clusters using the distance matrix. However, this method takes some time. In the Complete Linkage method, the merge process is performed by considering the largest distance between the data. The downside is that this method is sensitive to endpoints. The basic logic in the Average Linkage method is the process of applying merging by averaging the distances between the data. This method is more preferred. Ward's Linkage method aimed to minimize information loss. Information loss is the sum of the squares of the errors. This method does not combine groups with the smallest distance but does combine structures with the minimum value of variance. The Centroid Linkage method performs a grouping based on Euclidean distance. It is sufficient that the center points of the converging clusters are at a minimum distance from each other while performing the operation. In the Median Linkage method, the center of gravity of the larger data set is taken as the basis when merging the clusters, and the new center is closer to this center of gravity [18]. In the hierarchical clustering method, we can visualize with a dendrogram. A dendrogram is a tree-like structure describing the relationship between all data points in the system [19].

3.4. Model Building and Accuracy Measurement

While measuring the accuracy of the model, TP, FP, TN, and FN values were used. TP is the positive data, and the model says positive; FP is the negative data, but the model calls it positive; TN is the negative data, and the model calls it negative; FN represents data that is positive but predicted negatively by the model. These values were reached by comparing the outputs predicted by the model with the actual labels.

4. Results and Discussion

After the cleaning and editing studies on the data set, k-Means and Hierarchical clustering algorithms were applied. The test size was determined at a rate of 0.5 to be the same in the algorithms. Since there are 5 cancer types, the number of clusters was chosen as 5. After the necessary coding was done, the ARI score was 0.79 and the silhouette score was 0.135 in the k-Means algorithm (Table 1). The ARI score appears to be closer to the case of perfect duplication of values. Looking at the silhouette score, it is seen that the value is close to the average. Four linkage methods (full, ward, average, single) were tried in the hierarchical clustering algorithm and the best results were obtained in ward with ARI score and silhouette score of 0.775 and 0.13, respectively (Table 2). The reason why the silhouette score is low is that the ward method differentiates the classes from each other well. Looking at the Accuracy value, it was seen that accuracy was 0.990 for k-Means and 0.999 for Hierarchical clustering, slightly better accuracy than the k-Means algorithm (Table 3).

Table 1. ARI and Silhouette Score Results for k-Means

Adjusted Rand Score	0.7940732386133297
Silhouette Score	0.13507290311138614

Table 2. ARI and Silhouette Score Results for Hierarchical Clustering

	ARI	Silhoutte Score
complete	-0.030075	0.091312
ward	0.775358	0.133002
average	0.00041	0.254240
single	0.00041	0.201673

Table 3. Accuracy for k-Means and Hierarchical Clustering

	Accuracy
k-Means	0.9909194430904822
Hierarchical Clustering	0.9997240156180038

Then the dendrogram was drawn. The dendrogram obtained is as in Figure 1.

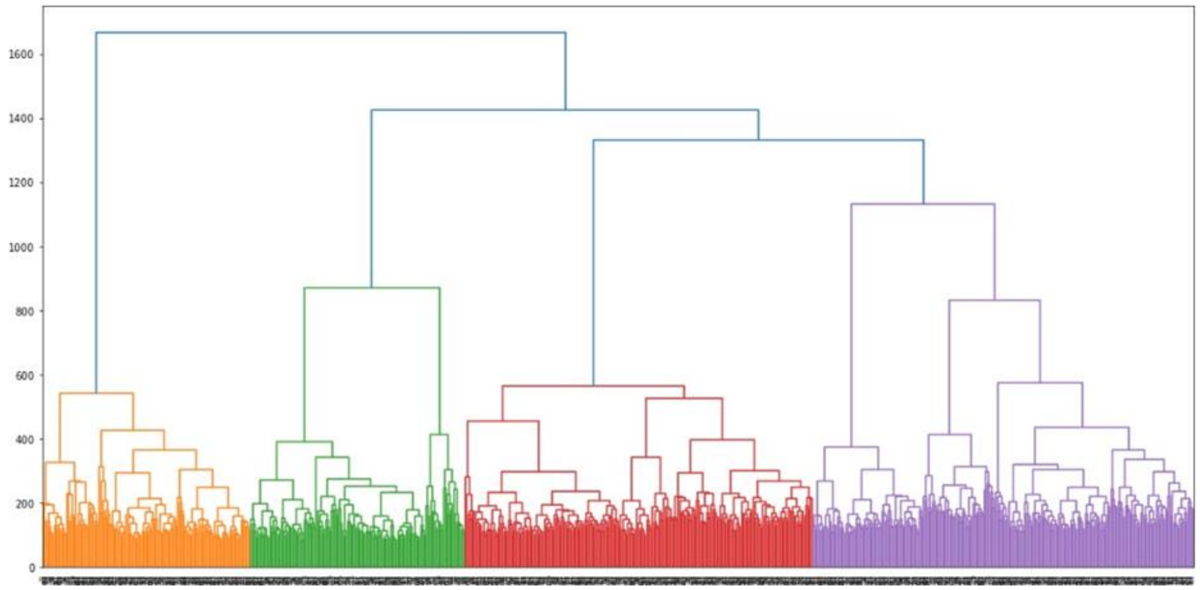

Figure 1. Dendrogram

In this study, two successful models with 99.0% and 99.9% accuracy values for cancer subtype prediction are presented. Since the algorithms are not deterministic algorithms, results can be obtained at once without the need to run the algorithm repeatedly. Considering the results, it is seen that this study is a reliable study that can be used for cancer prediction. Thanks to this study, it will be possible to determine whether a person has cancer yet at the gene level and to be diagnosed early. The patient will be able to get through this period with minimal wear, without the need for more drugs, chemotherapy, or radiotherapy. Thus, psychological, economic, and physical damage will be minimized. This study can be an informative resource for software that will be developed in the medical field and include data mining algorithms, and with this study, the contribution of clustering algorithms to the evaluation and analysis of various medical data can be observed. Compared with previous studies [20], this study showed a higher accuracy rate. It is seen that none of the algorithms used here reach 99% accuracy. The source codes of the study can be found at [21].

5. Conclusion

In this study, the performance of two different unsupervised machine learning methods was investigated on the data set containing the gene expression values of 5 cancer types (BRCA, KIRC, LUAD, LUSC, UCEC). K-Means and Hierarchical Clustering algorithms were used as clustering algorithms. The study was carried out on the Jupyter Notebook platform. Accuracy, adjusted rand score (ARI) and silhouette score metrics were used to measure the performance of the algorithms used. Values close to 1 for the ARI score indicate good clustering. Silhouette score values are -1 for very weak clusters, 1 for very dense clusters, and 0 for results where the clusters are well separated from each other. The study was visualized by drawing a dendrogram (Figure 1). K-Means and Hierarchical Clustering algorithms showed similar clustering performance. While the K-Means algorithm has an accuracy value of 99.09%, the Hierarchical Clustering algorithm has an accuracy value of 99.9%. Although both methods are successful, the Hierarchical Clustering algorithm has slightly higher accuracy. 4 different connection methods are used in the Hierarchical Clustering algorithm. These connection methods are complete, ward, average, and single. When the results were compared as in Table 2, it was seen that the "ward" method showed the best performance. The ARI score is 0.775 and the Silhouette score is 0.13. When these results are interpreted, the ARI score is close to 1. This result indicates a good clustering. Looking at the silhouette score, it is close to 0. This result shows that the clusters are well separated from each other. When the results of the K-Means algorithm are examined, it is seen that the ARI score is 0.79 and the silhouette score is 0.135. It also states that the k-Means algorithm creates successful clusters as in hierarchical clustering. Both clustering techniques used in this study showed successful results. Two models have been obtained that can be used for the early detection of cancer at the gene level. This study will shed light on future studies on the subject.

6. Acknowledgments

The authors are grateful to TUBITAK ULAKBIM High Performance and Grid Computing Center (TRUBA), Bursa Technical University High-Performance Computing Laboratory.

References

[1] Prat A, Pineda E,Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. Breast, 2015

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

[3] S. Saha, A. Ekbal, K. Gupta, S. Bandyopadhyay, Gene expression data clustering using a multiobjective symmetry based clustering technique, Comput. Biol. Med. 43 (11) (2013) 1965–1977, https://doi.org/10.1016/j.compbiomed.2013.07.021

[4] Fahad Hussain, Umair Saeed, Ghulam Muhammad, Noman Islam and Ghazala Shafi Sheikh, "Classifying cancer patients based on DNA sequences using machine learning", 2019

[5] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects", 2014

[6] KonstantinaKourou, Themis P.Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadisa, "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal, 2015

[7] Gunasekaran Manogaran, V. Vijayakumar R. Varatharajan, Priyan Malarvizhi Kumar, Revathi Sundarasekar, Ching-Hsien Hsu, "Machine Learning Based Big Data Processing Framework for Cancer Diagnosis Using Hidden Markov Model and GM Clustering", Wireless Personal Communications, 2018

[8] Zeid Khitan, Anna P. Shapiro, Preeya T. Shah, Juan R. Sanabria, Prasanna Santhanam, Komal Sodhi, Nader G. Abraham, and Joseph I. Shapiro, "Predicting Adverse Outcomes in Chronic Kidney Disease Using Machine Learning Methods: Data from the Modification of Diet in Renal Disease", Marshall Journal of Medicine, 2017

[9] Manish Motwani, Damini Dey, Daniel S. Berman, Guido Germano, Stephan Achenbach et al., "Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5- year multicentre prospective registry analysis", European Heart Journal, 2017

[10] Timor Kadir, Fergus Gleeson, "Lung cancer prediction using machine learning and advanced imaging Techniques", 2018

[11] Md. Milon Islam, Md. Rezwanul Haque, Hasib Iqbal, Md. Munirul Hasan, Mahmudul Hasan, Muhammad Nomani Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques", 2020

[12] Yixuan Li, Zixuan Chen, "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction", 2018

[13] Mingyang Lu, Zhenjiang Fand, Bin Xu, Lujun Chen, Xiao Zheng, Jundong Li, Taieb Znati, Qi Mi, Jingting Jiang, "Using machine learning to predict ovarian cancer", 2020

[14] UCI Machine Learning Repository, "Gene Expression Cancer RNA-Seq Data Set", 2016, <https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq>, (24 May 2022)

[15] Laiqa Rukhsar , Waqas Haider Bangyal , Muhammad Sadiq Ali Khan , Ag Asri Ag Ibrahim, Kashif Nisar and Danda B. Rawat, "Analyzing RNA-Seq Gene Expression Data Using Deep Learning Approaches for Cancer Classification", 2021

[16] Imad Dabbura, "K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks" 2018, <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks- aa03e644b48a>, (01 June 2022)

[17] Dr. Michael J. Garbade, "Understanding K-Means Clustering in Machine Learning" 2018, <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>, (01 June 2022)

[18] Eyup Kaan Ulgen, "Hierarchical Clustering", 2021, <https://www.veribilimiokulu.com/hiyerarsik-kumeleme/>, (15 June 2022)

[19] Prasad Pai, Hierarchical Clustering Explained, 2021, <https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8>, (10 June 2022)

[20] Fahad Hussain, Umair Saeed, Ghulam Muhammad, Noman Islam and Ghazala Shafi Sheikh, "Classifying cancer patients based on DNA sequences using machine learning", 2019

[21] Seyma Dogru, "Predicting Cancer Using Machine Learning on DNA Sequences", 2022, <https://github.com/seymadogru/Prediction-of-Cancer-in-DNA-Sequences-Using-Unsupervised-Learning-Methods-.git >, (03 September 2022)