# Identification of Gene of Melanoma Skin Cancer Using Clustering Algorithms

Mohanavali Sithambranathan [a], Shahreen Kasim [a,1,*], Muhammad Zaki Hassan [a], Nur Aniq Syafiq Rodzuan [a]

[a] Faculty of Computer Science and Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia.
[1] shahreen@uthm.edu.my
* corresponding author

ARTICLE INFO

**Article history**
Received April 28, 2020
Revised May 08, 2020
Accepted May 11, 2020

**Keywords:**
melanoma
skin cancer
identification
gene
clustering algorithms

ABSTRACT

The Melanoma is the deadliest skin cancer. It can be developed in any parts of the human body. The cancer disease can be cured if it is diagnosed early and proper treatment is taken. In cancer classification, there is a problem in handling the large data of cancer. Large data contains meaningless data and redundant data. Therefore, to overcome the problem, many computer approaches for classification have been proposed in the previous literature. This time, the clustering process for melanoma is conducted using Support Vector Machine and K-Means. Therefore, the purpose of this research is to identify and evaluate the performance of the accuracy of genes that contain melanoma skin cancer using the clustering algorithms.



## 1. Introduction

Cancer is an abnormal growth of cell (Louise, 2018). There are many types of cancer such as breast cancer, lung cancer, skin cancer and colon cancer. Cancer can be cured if it is diagnosed early and proper treatment is taken. The skin is an organ that separates human body from the environment. Due to that, skin cancer becomes an ordinary type of cancer that affects humans. The number of cases of skin cancer is increasing day by day. Researchers found that the United Kingdom (U.K) is the fast rising country when it comes to skin cancer patients. The two types of skin cancer that exist are melanoma and non-melanoma skin cancer (Wilson, (2012). Melanoma is a type of cancer that develops from the pigment-containing cells known as melanocyte (Talantov, Mazumder and Jack, 2005). Non-melanoma skin cancer refers to all the non-melanoma types of cancer that occur in the skin (Wolters Kluwer, 2019). Skin cancers occurs due exposure to sun and also affected from the genetic problem (Moore, 2001).

In the previous literature, the problem in classifying a cancer is when the gene expression data is used (Lu and Han, 2003). Gene expression data is considered a high-dimensional type of data. Therefore, the analysis of gene expression is difficult to conduct because of the big data that contains noise, redundant data, and unrelated information of features.

In this research, K-Means and Support Vector Machine were used to cluster and classify data and compare the detection accuracy. Clustering involves assigning data points to a cluster where items in the same cluster are the same. Therefore, the clusters are known by some similarity measures for example distance, connectivity and intensity (Narongsak and Anongnart, 2016, Bernhard, 2001).

## 2. Literature Review

This section discusses the finding of literature reviews related to the research.

### 2.1. Support Vector Machine

A supervised learning system is the Support Vector Machine (SVM) George et al., 2011. SVM based algorithms used for identification and regression processing to analyze data and understand trends. An algorithm for SVM learning creates a prototype that assigns new examples to one or the other group, rendering it a non-probabilistic conditional linear classifier.

George et al., 2011 stated that SVM is the best cancer classification system to use. There are several explanations why in cancer classification, SVM has the best performance. Next, SVMs have checked the potential not only to correctly classify organizations into relevant categories, but also to distinguish situations where the evidence does not help understand the classification. SVM has many computational features that make them interesting in the study of gene expression, including their stability in selecting a similarity variable, the lack of solution while dealing with large data sets, the ability to handle wide field spaces, and the ability to identify outliers. To construct a classifier the following formula is used.

$$y\,(x) = \text{sign}\left[\sum_{k=1}^{N} {}_{k}\, y_k\, \psi\,(x, x_k) + b\right] \quad (1)$$

This formula consists of real constant and polynomial SVM degree

### 2.2. K-Means

K-means is an incremental clustering method which dynamically integrates one cluster center at a time by way of a deterministic global search process consisting of N (with N being the width of the data set) executions of the k-means algorithm from correct initial positions (Lozano et al., 1999). The K-means method in data mining begins with a first group of randomly selected centroids, which is used as the starting points for each cluster, and then conducts iterative (repetitive) calculations to refine centroid (*ci* and *cj*, *ci* ≠ *cj*) locations. The new means (centroids) of the observations are then calculated in the new clusters.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{z_j C S_i^{(t)}} x_j \quad (2)$$

The calculation shows that the algorithm has converged when the assignments no longer change. The algorithm does not guarantee that the optimum can be found. The algorithm is often presented as assigning objects to the nearest cluster by distance.

## 3. Methodology

In the first phase which is data pre-processed, the data is collected from GEO database. Melanoma skin cancer gene expression (GSE3189) is obtained from the GEO database. Affymetrix Human Genome U133A Array is the basis of the software used. GSE3189 contains three types of classes namely normal skin, nevi skin and melanoma (Lu and Han, 2003). There are 70 samples in this set of data. 7 are normal skin, 18 are nevi skin and the remaining 45 are melanoma. Next, pre-processed data is used in the clustering and classifying process where it uses the Support Vector Machine (SVM) and K-Means. In phase three, the documentation of paper works, soft materials, and coding design used in the research are prepared in the form of paper works. The purpose of documentation is to provide a clear understanding to the readers about the overall flow of the research.

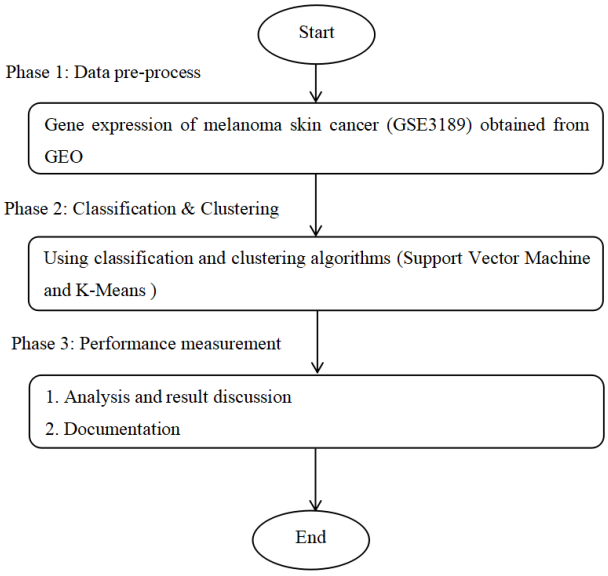


**Fig 1.** The overall flow of research methodology

For SVM the first step is to import the dataset into the R environment. Then, the specific code that describes the SVM function to plot the graph is used. The gene expression and selected genes are used as the input and parameter for the SVM. The results will be produced in the format of graph.

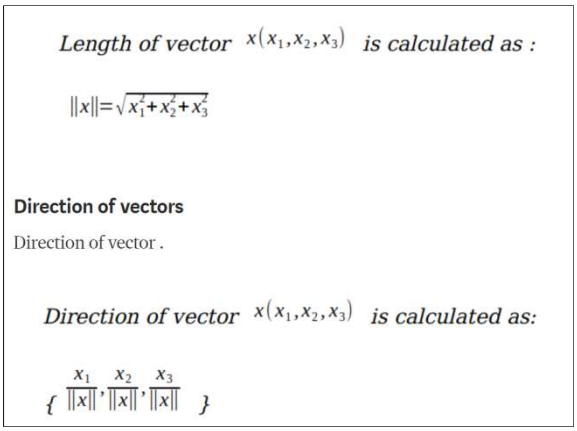


**Fig 2.** This is a mathematical equation used to calculate the length and direction of vector

As for K-Means the same method is used where first of all the dataset needs to be imported into the R environment. Next, the K-Means function is used to cluster the data of genes into groups. The results will be produced in 2D representation where the genes will be grouped according to the type of skin. The algorithm works as follows:

- Step 1: Choose groups in the feature plan randomly.
- Step 2: Minimize the distance between the cluster center and the different observations (centroid). It results in groups with observations.
- Step 3: Shift the initial centroid to the mean of the coordinates within a group.
- Step 4: Minimize the distance according to the new centroids. New boundaries are created. Thus, observations will move from one group to another

Repeat until no changes are observed in groups.

K-means usually takes the Euclidean distance between two features :

$$distance\ (x, y) = \sum_i^n (x_i - y_i)^2 \quad (3)$$

Different measures are available such as the Manhattan distance or Minlowski distance. It is noted that K-mean returns different groups each time you run the algorithm. We recall that the first initial guesses are random and the distances are computed until the algorithm reaches a homogeneity within groups. This means that k-mean is very sensitive to the first choice, and unless the number of observations and groups is small, it is almost impossible to get the same clustering.

## 4. Result and Discussion

There are two types of algorithms used to conduct this research which is SVM and K-Means. SVM is the supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. K-means is simple where it groups similar data points together and discover underlying patterns. Figure 3 and 4 show the results for the classifiers' performance.


**Fig 3.** Result of normal and melanoma for GSE3819/206403 using Support Vector Machine

Figure 3 shows the results of SVM. SVM is a subclass of supervised classifiers that attempt to partition a feature space into two or more groups. The separation boundary is linear, leading to groups that are split up by lines (or planes) in high-dimensional spaces. *y* as the response variable and other variables serve as the predictors. The data frame will have unpacked the matrix *x* into 2 columns named *x1* and *x2*. Based on the result, the number of support vectors is 6 and they are the points that are close to the boundary or on the wrong side of the boundary. The support vector 6 is the number of o in the graph. The blue color part indicates the melanoma skin genes while the pink color indicates the normal and nevi skin genes. The points that are close to the boundary are colored blue while the wrong side boundary is pink. The wrong side of the boundary shows that the observed data is sufficiently inconsistent. This shows that the dataset is not grouped properly as the dataset is mixed up as shown in the graph.

Based on figure 4, to perform the analysis, two groups of skin that contain different genes were selected. The data is from the same source and it is tested using k-means algorithm. According to the result the genes are clustered into two groups. It explains the point variability where a centroid is the imaginary or real location representing the center of the cluster. The medium size grouped

data is nevi and the largest grouped data is melanoma skin. A cluster refers to a collection of data points aggregated together because of certain similarities. Therefore, by referring to the result in figure 3, one can conclude that the highest number of centroid in one place is considered as normal cell. According to the datasets there are 45 genes that are melanoma while 18 are nevi. Therefore, the clustering process was carried out to identify groups of similar genes and to group the genes according to the group to which they belong.

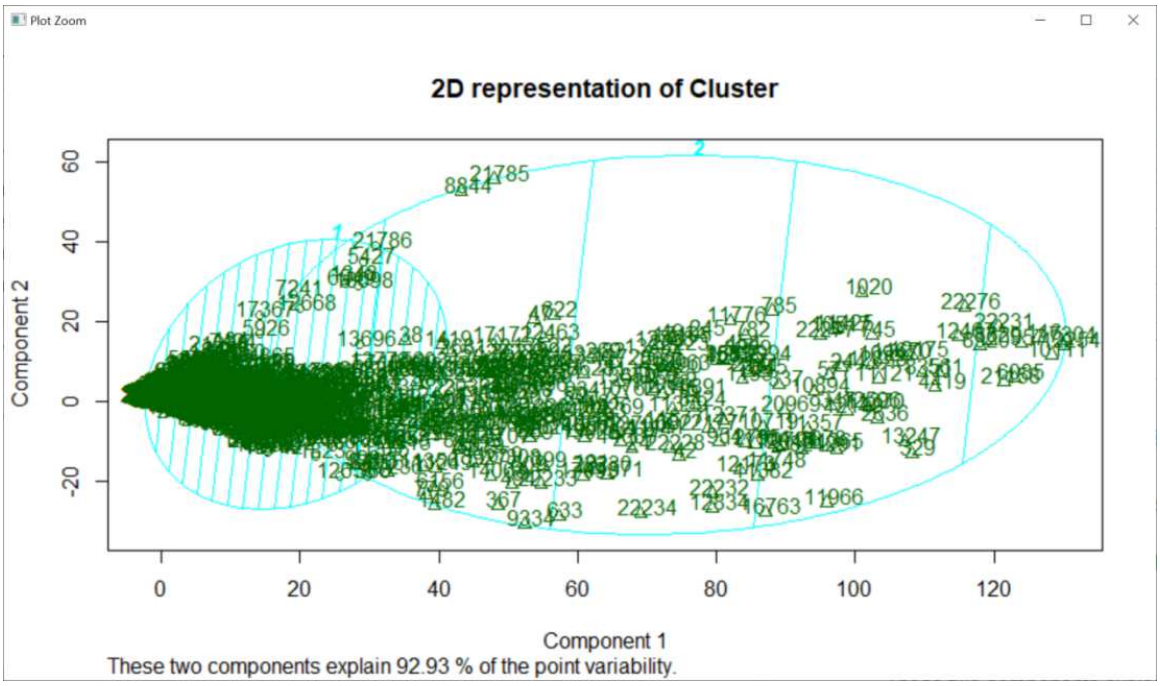


**Fig 4.** Result of 2D representation for GSE3819 using K-Means

Thus, the overall results obtained from both SVM and K-Means show that the graphs produced by both algorithms are different from one another. According to the result of the SVM graph, there are 2 boundaries that consist of points which are *x* and *o*. If the value of *X2* is 0.01 then the value below it is considered as the data that is sufficiently inconsistent. As for the K-means graph the points which are in the number form are grouped into two labeled as *1* and *2*. The largest grouped data is known as melanoma skin according to the raw dataset calculation.

## 5. Conclusion

As a conclusion, this research focuses on identifying the result of clustering using K-Means and Support Vector Machine. It is to determine the high accuracy of clustering towards the melanoma data. From the clustering process, the genes will be identified which are most related to the melanoma skin cancer. In the process of classifying the cancer, the data sets of genes expression will be collected. To get the most suitable data for this research, the pre-processed data sets will be done first. The best genes will be obtained through the process of selecting the features using limma package that is used to calculate and normalize the data. Other than that, the package also focuses on differential expressed genes analysis. To find out the accuracy of the selected genes, a classifier will be constructed once the pre-processed data sets are done. Based on the results, K-Means is good at grouping dataset compared to SVM. This is because the results obtained from the K-Means are grouped into two according to their criteria but for SVM the results were mixed.

## Acknowledgment

I would like to express my special thanks to my supervisor Prof. Madya Dr. Shahreen Binti Kasim for the support and guidance in helping me throughout the course of my research and for simply not giving up on me.

## References

[1] Bernhard Scholkopf, Alexander J. Smola. (2001). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.

[2] Victo Sudha George, Cyril Raj. Review On Feature Selection Techniques And The Impact Of Svm For Cancer Classification Using Gene Expression Profile. International Journal of Computer Science & Engineering Survey [Internet]. Academy and Industry Research Collaboration Center (AIRCC); 2011 Aug 30;2(3):16–27. Available from: http://dx.doi.org/10.5121/ijcses.2011.2302

[3] Peña J., Lozano J., Larrañaga P. An empirical comparison of four initialization methods for the K-Means algorithm. Pattern Recognition Letters [Internet]. Elsevier BV; 1999 Oct;20(10):1027–40. Available from: http://dx.doi.org/10.1016/s0167-8655(99)00069-0

[4] Mulryan C. Understanding cancer: the basics. British Journal of Healthcare Assistants [Internet]. Mark Allen Group; 2010 Jun;4(6):266–9. Available from: http://dx.doi.org/10.12968/bjha.2010.4.6.48484

[5] Lu, Y., & Han, J. (2003). Cancer classification using gene expression data. Information System, 28(4), 243-368.

[6] Hang X. Cancer classification by sparse representation using microarray gene expression data. 2008 IEEE International Conference on Bioinformatics and Biomeidcine Workshops [Internet]. IEEE; 2008 Nov; Available from: http://dx.doi.org/10.1109/bibmw.2008.4686232

[7] Georgia Moore. A course of action on skin cancer. Business and Health 2001;4:40.

[8] Chayangkoon N, Srivihok A. Two Step Clustering Model for K-Means Algorithm. Proceedings of the Fifth International Conference on Network, Communication and Computing - ICNCC '16 [Internet]. ACM Press; 2016; Available from: http://dx.doi.org/10.1145/3033288.3033347

[9] Talantov D. Novel Genes Associated with Malignant Melanoma but not Benign Melanocytic Lesions. Clinical Cancer Research [Internet]. American Association for Cancer Research (AACR); 2005 Oct 15;11(20):7234–42. Available from: http://dx.doi.org/10.1158/1078-0432.ccr-05-0683

[10] Wilson MA, Nathanson KL. Molecular Testing in Melanoma. The Cancer Journal [Internet]. Ovid Technologies (Wolters Kluwer Health); 2012;18(2):117–23. Available from: http://dx.doi.org/10.1097/ppo.0b013e31824f11bf

[11] Corner C, Hoskin P. Skin cancer. Oxford Medicine Online [Internet]. Oxford University Press; 2013 May; Available from: http://dx.doi.org/10.1093/med/9780199696567.003.0018