GRAPHICAL MODEL AND CLUSTERING-REGRESSION BASED METHODS FOR CAUSAL INTERACTIONS: BREAST CANCER CASE STUDY

Suhilah Alkhalifah and Adel Aloraini

Computer Science Department College of Computer, Qassim University, Saudi Arabia

ABSTRACT

The early detection of Breast Cancer, the deadly disease that mostly affects women is extremely complex because it requires various features of the cell type. Therefore, the efficient approach to diagnosing Breast Cancer at the early stage was to apply artificial intelligence where machines are simulated with intelligence and programmed to think and act like a human. This allows machines to passively learn and find a pattern, which can be used later to detect any new changes that may occur. In general, machine learning is quite useful particularly in the medical field, which depends on complex genomic measurements such as microarray technique and would increase the accuracy and precision of results. With this technology, doctors can easily diagnose patients with cancer quickly and apply the proper treatment in a timely manner. Therefore, the goal of this paper is to address and propose a robust Breast Cancer diagnostic system using complex genomic analysis via microarray technology. The system will combine two machine learning methods, K-means cluster, and linear regression.

KEYWORDS

Microarray, K-means cluster, and linear regression.

1. INTRODUCTION

During the last decades, cancer has been one of the main focus and concern fields for scientists. Cancer starts in cells which are the basic structural unit in the body. Cells of the cancer patient are multiplying in a way that is difficult to control, those cells cause a tumor which can be benign or premalignant. Benign tumors are usually not harmful and do not spread to other parts of the body, which is the opposite of cancer lumps. Breast Cancer is one of the most types of cancers which infects women, that is because of changes in lifestyle, increased age, and hormonal disorders [1]. To study every gene in a cell, scientists spend a lot of time when they use manual tools to monitor gene's behavior.

Alternatively, they use gene expression microarray technology to study complex relations between various genes in genomics, thousands of genes or even every gene in an organism all at once, [2]. The microarray technology has been widely applied to the range of machine learning methods mainly to understand how data are obtained and analyzed [3]. Machine learning is a branch of artificial intelligence and is an effective technique to classify data, it mostly used in medical fields to diagnosis and treatment [4]. Moreover, Machine learning can help doctors take accurate decisions when they discover the relationship between datasets from prior cases [5].

In our study, we use a combination of two machine learning methods, linear regression, and k-means. K-means is mainly used to find groups based on similarity and dissimilarity in the data while the main purpose of using linear regression is to predict the amount of relationship between variables obtained from the set of data [6]. In addition to its simplicity and capacity to interpret the coefficients as interactions in the underlying network [7]. This study will be as a contribution to what exists in the literature for feature selection research, as well.

The rest of the paper is organized as follows: Section 2 presents an overview of DNA-Microarray and reviewed several previous studies that use machine learning to microarray data analysis. Section 3 provides a description of the methods and software utilized in this study. Section 4 shows the different results found. For each result, we provide scientific reasons from our perspectives on whether we accept the result or not. Section 5 concludes the paper.

2. RELATED WORK

Microarray technology has rapidly established a critical site in cancer research since 1996 [8]. The capacity of microarrays becomes the main attraction for molecular biologists. By observation of the expressional behavior of the genes under various experiments [9]. In addition, a microarray can be applied in a variety of fields such as drug discovery, genetics, and microbiology [10]. The purpose of using a microarray differs from one researcher to another. For example, the researcher may interest in finding the small change in gene expression that affects phenotype. Another researcher may interest in understanding the architecture of genetic from the whole gene expression that describes the mechanisms of transcription control in a cell [11]. The accurate result depends on the careful design of the experiment and the appropriate selection for the machine learning methods to extract the relationships between genes [12].

Through machine learning methods, it is possible to obtain an early diagnosis of cancer disease from data that is collected by microarray gene expression technology [13]. Machine learning methods can help doctors to make the correct decisions from discovering the relationship between cases and predict the disease results using similar previous cases [5]. There are many types of machine learning methods that can be applied to analyzing big data. For example, in [14], the authors review two types of machine learning methods, which are Decision Trees and Artificial Neural Networks. Using these methods can explore microarray data in a quick way. The result was Decision Trees outperformed Artificial Neural Networks, because of Decision Trees can inspect the result by humans directly. In a study [15], compared the performance of several classification methods on different microarray datasets. Classification methods include Support Vector Machine, Decision Tree, Radial basis function Neural Nets, Bayesian, Multi-layer perceptron Neural Nets, and Random Forrest. The results reported the importance of accurate classification was related to feature selection and the number of genes and samples. As well in [16], the authors compared different classification methods which are Support Vector Machine, Structure Adaptive Self-organizing map, K-Nearest Neighbor, and Multi-Layer Perceptron. The results indicate a link between classifiers and features, which act as a guide to choosing the best method for bioinformatics problems.

In our study, we focused on researches that used linear regression and K-means methods to analyze microarray data. We have obtained a set of research that used one of these methods without combined them for example, in [17], the authors obtain prior knowledge of genes from the Gene Ontology database to improve the regression model. This model assumes the similarity of weights between genes to reduce the errors of regression that arise from the spread data. [2] is another study about microarray data analysis, where the authors improve a new method called linear regression-based feature selection to obtain for accurate classification of a dataset by using fewer features. In [18], using K-means clustering to compare various types of initialization value

and distance to classify the patients depend on the attributes of breast cancer. There is another study in [19], the authors evaluated different types of clustering algorithms to help scientists to get the best group of a gene. The best algorithm was Diana that gave the best number of clusters. In [20], applied K-means and K-nearest-neighbor on data of gene expression to predicting breast cancer survival. The fitting calibration slope gave the best outcome. In a study [21], the authors presented a new method to avoid the error in a number of clusters. This method applied to yeast data and it gave a great result.

There are two important objectives for gene expression data which are the identification of differentially expressed genes and clustering of genes. With linear regression, we can predict the relationship between gene expression data [22], and with the clustering method we can find out biological meaningful groups. Such groups are helpful for further studies including gene function and regulation [9]. However, there are only limited related studies have been combined between linear regression and K-means clustering. To the best of our knowledge, there are only two studies [23], [24]. Those studies used the same method, clustering of the regression model, with a different dataset. This method applies regression on gene expression then assumes gene clusters depending on regression coefficients similarity. It can apply in a complex experiment and it will provide an accurate cluster. However, K-means is more useful to examine the data that do not have prior knowledge in correlations between genes such as in our case.

3. METHODOLOGY

A. K-means Clustering

K-means clustering is an unsupervised learning algorithm mostly used with unlabeled data to solve clustering problems. It can be defined as the task of finding subgroups in datasets that are very similar while data in other clusters are very different.

K-means clustering follows a simple procedure to classify any dataset into a number of clusters by initially selecting the number of clusters and based on this the centroids are randomly set. Each data point is assigned to the nearest centroid. The centroid is updated based on the data points in the cluster until it reaches a limit when there is no change in the cluster's data [25]. In the following, we provide details on how the k-means algorithm works.

***K-means* algorithm**

Given: data {x} = {$x_1, x_2, \dots, x_n$}, number of cluster {k}
Where: x represents a gene

BEGIN initialization
initialize the cluster center randomly v_i
END initialization

BEGIN iteration
Stop = false
while Stop = false do
for (i = 1; i ≥ c_i ; i ++) do

$$\text{Euclidean distance} = \sqrt{(x_{i1} - v_{j1})^2 + (x_{i2} - v_{j2})^2 + \cdots + (x_{in} - v_{jn})^2}$$

to assign x_i to the closest v_i
end for
for (j = 1; j ≥ k ; j ++) do

$$v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_i$$

end for
if
Stop = true
end if
end while
END iteration

As we see the steps to implement the algorithm are:

1. The number of clustering is given (k).
2. Initialization step:
   i) Randomly select the cluster center (vj)
3. Iteration step:
   i) Calculate the distance between the centers and each point using Euclidean distance.

Where ||xi - vj|| is the Euclidean distance between xi (the set of data points) and vj (the set of centers), ci is the number of data points in ith cluster, c is the number of cluster centers.

   ii) Assign the data point to the nearest cluster center.
   iii) Recalculate the new cluster center using vj.
   iv) Repeat until no data point was reassigned.

B. Linear Regression

Linear Regression is a method for predicting the relationship between variables. The value of the dependent variable is based on the given independent variable. It is relatively considered a simple and most useful algorithm. There are different types of linear regressions, simple linear regression and multiple linear regression are some of them. Throughout this study, we focus on multiple linear regression. The equation of multiple linear regression is: $Y=\beta_0+\beta_1x_1+ \beta_2x_2+\varepsilon$ [26].

In this study:

- Y represents the effect variable.
- Xs represents the causal variables.
- β represents the coefficient value. The coefficient value represents the amount of the correlation e.g. if the coefficient value is high that means there is a high correlation between these two genes and vice versa.
- The sign represents the type of correlation. e.g. if the sign is positive that means there is a positive correlation between those two genes, in another word, if the value of one gene is increased the value of another gene will be increased as well and conversely in the inverse correlation, if the value of one gene is increased the value of another gene will be decreased.

C. WEKA Software

WEKA stands for Waikato Environment for Knowledge Analysis, and it was developed at Waikato University in New Zealand. It is open-source software written in Java and used for data mining tasks. WEKA includes a library of the machine learning algorithm, such as clustering, regression, and classification [27]. The main advantages of using WEKA in this research is that it is freely available, easy to use by providing a user-friendly GUI, and has a large collection of data preprocessing and modeling methods [28]. The WEKA version used in this study is version 3.8.2 installed on Windows 7.

D. Dataset

The dataset, that contains 209 Breast Cancer samples were obtained from [29] via microarray technology. Due to the limitation of our hardware, especially in WEKA memory. We just included 30 samples. Fig. 1 shows our dataset, in which the first column represents the name of genes, and the rest of the columns represent gene expressions.

MAPK3	7.9	8.11	8.07	7.95	8.18	7.93	7.58	7.12	6.97	8.15	7.57	8.31	7.8	7.88	8.26	8.46	8.31	7.76	8.03	8.19	7.43	8.3	8.04	8.34
DUSP1	6.21	5.81	5.8	6.28	6.03	6.7	6.08	5.84	5.83	6.54	6.25	6.65	6.23	6.34	5.54	5.73	5.44	5.57	6.22	5.97	6.35	6.54	5.67	6.05
MAPK11	7.03	6.63	6.855	7.375	6.815	7.26	6.715	6.71	6.675	6.775	6.88	7.155	6.62	6.915	6.07	6.52	6.295	6.26	6.445	6.41	6.37	6.475	6.355	6.43
RELA	6.8625	6.9075	6.82	7.1225	7.1975	6.3775	6.465	6.015	6.1775	6.905	7.01	6.435	6.655	6.4725	6.36	6.95	6.6425	6.0475	6.6625	6.6175	6.715	6.945	6.5225	6.58
IL1A	3.99	3.83	3.84	3.87	3.45	4.03	3.78	3.8	3.93	3.68	3.77	3.48	3.74	3.53	3.28	3.11	3.29	3.36	3.64	3.2	3.23	2.92	3.21	3.12
BDNF	3.485	3.335	4.005	3.45	3.43	3.56	3.52	3.32	3.27	3.4	3.465	3.295	3.42	3.585	2.875	2.865	2.865	3.19	2.99	2.97	2.71	2.995	2.8	2.855
PDGFA	7.015	6.81	6.78	6.815	6.995	7.215	6.91	6.77	6.925	6.66	6.51	6.84	6.73	6.74	5.81	6.33	6.095	5.825	6.69	5.61	6.095	6.36	6.13	6.36
RPS6KA	8.32	8.3	7.92	8.67	8.15	7.94	7.74	7.69	7.78	8.26	7.84	8.08	7.79	7.76	8.21	8.17	8.15	7.14	8.37	7.99	8.66	8.17	8.24	7.85
MAP2K1	6.14	6.26	6.365	6.475	6.47	6.3	6.255	6.525	6.68	7.105	6.32	6.61	6.43	6.615	6.57	6.79	6.575	6.84	6.61	6.725	6.85	6.965	6.825	7.025
MAP2K2	6.93	6.71	6.805	7.315	6.71	7.155	6.77	6.37	6.45	6.745	6.795	6.88	6.875	6.475	7	6.845	6.815	6.84	7.305	6.94	6.66	7.005	6.875	6.835
MAPT	6.006	5.682	5.702	6.504	5.488	6.832	6.086	5.724	5.542	5.39	5.966	6.44	6.16	6.42	5.416	5.196	5.112	5.838	5.398	5.828	5.048	5.794	5.852	5.45
FGFR2	4.86	4.395	4.5633	5.0633	4.53	5.2517	4.8517	4.6617	4.8117	4.585	4.71	5.07767	5.1433	4.9067	5.2133	4.6867	4.4533	5.05	4.73	5.2433	5.0517	5.835	5.1733	4.791
MOS	4.45	4.07	4.06	4.6	4.25	4.71	4.09	4.08	4.1	3.97	4.37	4.22	4.05	4.51	4.06	3.92	3.96	4.04	4.18	4.16	3.84	3.87	4.11	4.14
TRAF6	5.25	5.26	5.09	5.14	5.13	5.07	4.89	5.21	5.38	5.1	5.23	4.98	4.96	4.87	5.27	5.39	5.22	5.31	4.97	5.13	5.33	5.24	4.93	5
PRKCB	4.8975	4.54	4.4675	4.845	4.2525	4.7925	4.63	4.67	4.5575	4.6175	4.5325	4.5875	4.51	4.575	3.805	3.9975	3.945	4.0875	4.01	3.905	3.8375	3.915	3.75	4.0675
NF1	4.4833	4.3433	4.3833	4.7183	4.2517	4.7767	4.6417	4.4733	4.3567	4.2633	4.3833	4.505	4.4083	4.6267	3.79	3.815	3.7217	4.125	3.865	3.7967	3.665	3.9533	3.7767	3.881
MAPK9	6.02	7.465	7.045	6.13	7.53	6.49	6.83	8.05	7.39	6.65	7.685	6.585	7.78	7.43	7.56	7.565	7.14	7.74	6.58	6.15	7.245	6.79	7.775	6.815
PAK2	5.2925	4.965	4.9875	5.2275	5.45	5.025	5.3425	5.285	5.04	4.9475	5.1775	5.2125	5.2925	5.315	4.61	4.66	4.6225	4.91	4.1125	4.2525	4.635	4.73	4.4525	4.24
TGFB2	4.57	4.195	4.17	4.59	4.395	4.81	4.4	4.375	4.19	4.2	4.935	4.5	4.73	4.33	3.67	3.505	3.36	3.925	3.505	3.35	4.04	3.325	3.38	3.72
MKNK2	6.665	6.02	6.02	6.83	5.805	7.475	6.77	6.22	5.965	6	6.35	7.075	7.15	7.08	5.37	5.78	5.68	6.955	5.505	6.215	5.41	6.19	5.63	6.23
MAP3K13	4.12	3.805	3.96	4.345	4.01	4.34	4.065	3.725	3.705	3.915	3.86	4.115	4.09	4.045	3.405	3.445	3.52	3.765	3.76	3.665	3.35	3.615	3.615	3.9
FGFR4	6.71	6.4	6.41	6.92	6.16	6.71	6.06	6.23	6.18	6.33	6.28	6.47	5.95	6.11	5.61	6.32	5.85	5.64	6.06	6.08	5.47	6.17	5.67	5.87
DUSP2	6.1	5.8	5.98	6	6.35	5.98	5.91	5.76	5.72	6.13	5.83	6.25	5.95	6.02	5.55	5.59	5.54	5.12	5.59	5.95	5.55	5.83	5.49	5.98
MAP3K5	5.59	6.61	5.03	5.72	5.1	4.51	4.45	5.33	5.73	5.07	4.9	3.83	4.01	3.76	4.52	5.24	6.18	4.68	5.5	3.84	5.99	3.67	3.98	5
MAP3K3	6.58	6.45	6.33	6.81	6.37	6.6	6.04	5.84	6.21	6.25	6.08	6.26	6.39	6.33	5.88	5.98	5.82	5.9	6.4	6.16	5.77	5.93	5.94	6.05
CDC25B	8.51	8.45	10	8.16	9.3	6.82	7.13	6.95	7.65	8.1	7.38	7.64	7.35	7.46	8.72	8.26	7.87	7.77	8.94	9.27	8.24	8.21	8.51	8.62
NTRK2	4.394	4.278	4.138	4.638	4.296	4.802	4.46	4.434	4.552	4.298	4.666	4.506	4.216	4.306	3.554	3.834	3.902	4.13	4.672	3.706	4.2	3.686	3.724	4.302
IL1R1	7.86	9.62	9.38	7.78	7.12	8.48	6.83	8.42	10.12	9.63	8.94	8.11	7.38	6.45	7.4	8.78	8.14	9.03	8.13	5.99	9.5	6.44	7.24	8.1
NFKB1	7.2333	7.35	7.2233	7.3367	7.3967	7.47	7.3	7.56	7.9033	8.0567	7.4933	7.6633	7.97	7.5167	8.0667	7.94	7.8267	7.6167	8.1267	7.9667	7.9267	7.7433	7.9133	7.8533
FGF7	4.1767	3.76	3.6967	3.9167	3.6933	3.8433	3.8133	3.6167	3.7	3.82	3.62	3.5467	3.53	3.7167	3.35	3.3233	3.26	3.4333	3.77	3.2367	3.7667	3.2833	3.1233	3.36

Fig.1 Dataset

4. RESULT

A. K-means Experiments

Initially, we applied K-means algorithm to find groups based on similarity and dissimilarity in the dataset. To do this work we applied some experiments until we got a satisfactory result.

First, we start with the all gene in the dataset (30 genes) and we use two clusters to found the largest number of associated genes. The result was: (MAPK3- DUSP1-MAPK11 - RELA - PDGFA - RPS6KA1 - MAP2K1 - MAP2K2 - MAPT - MAPK9 - MKNK2 - FGFR4 - DUSP2 - MAP3K3 - CDC25B - IL1R1 - NFKB1) in cluster one, and (IL1A - BDNF - FGFR2 - MOS - TRAF6 - PRKCB - NF1 - PAK2 - TGFB2 - MAP3K13 - MAP3K5 - NTRK2 - FGF7) in cluster two. We remove MAPK3 and IL1A, which represents a class of gene name in WEKA, because we have identified the genes associated with them. Then we repeat the same operation for all remained genes.

Second, we applied the same previous experiment with a minor change, in which we removed one gene at a time from the most numerous clusters. This is the most efficient approach, to get an accurate result from the previously discussed results, because we give more chance for each gene.
From the first and second experiments, gene RPS6KA1 for example associated with (MAP2K1 - MAP2K2 - MAPK9 - MKNK2 - FGFR4 - MAP3K3 - CDC25B - IL1R1 - NFKB1) in the first experiment, while the same gene associated with (MAP2K1 - MAP2K2 - MAPT - MAPK9 - MKNK2 - FGFR4 - DUSP2 - MAP3K3 - CDC25B - IL1R1 - NFKB1) in the second experiment. We observed a difference in the genes associated with it, such as there is an association with MAPT and DUSP2 in the second experiment but it is not shown in the first experiment. Another example is NF1 associated with (PAK2 - TGFB2 - MAP3K13 - MAP3K5 - NTRK2 - FGF7) in the first experiment, while the same gene associated with (PAK2 - TGFB2 - MKNK2 - MAP3K13 - FGFR4 - DUSP2 - MAP3K5 - MAP3K3 - NTRK2 - FGF7) in the second experiment. There is an association with MKNK2, FGFR4, DUSP2 and MAP3K3 in the second experiment but it is not existing in the first experiment. When we realized different results, we decided to try an alternative operation due to the following reasons:

- When we remove one or two genes from the dataset, the number of associations for the remaining genes is reduced.
- In each step, we set two clusters. However, this does not mean that the two clusters are an optimal number of clusters. This obviously happens due to a weakness in the K-means method.

In the following, we started from the largest number of clusters, i.e. 29 clusters to the least number of clusters, i.e. 2 clusters and we do not remove any gene from the dataset. In each step, we extracted the genes that have more associated with each other. For example, when we specified the number of clusters to 29, we observed MOS and NF1 belong to the same cluster (which is cluster number 12) and each other gene in a separate cluster, as shown in Fig. 2.

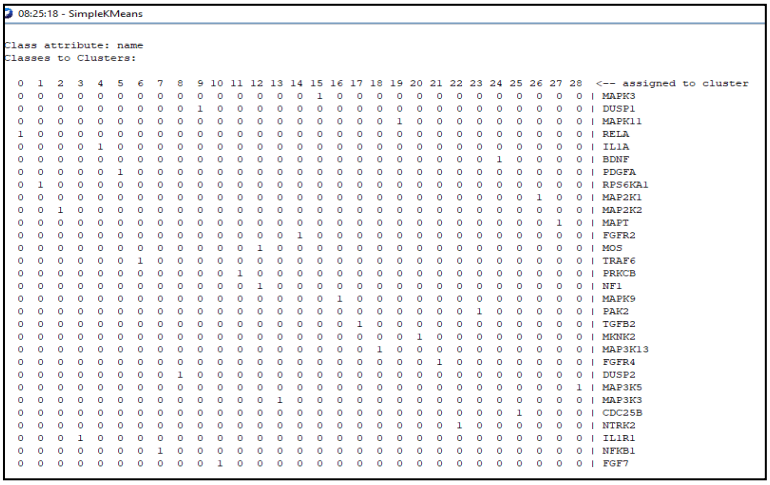

Fig.2 A Result of Cluster Number 29

We Applied The Same Operation For All Possible Number Of Clusters And We Get The Result Shown In Table 2.

Table 2. A Result Of The Third Experiment Using K-Means

Cluster Number	**Cluster Result**	**Associations**
29	Cluster 1	**MOS - NF1**
28	Cluster 1	MOS - NF1
	Cluster 2	**TRAF6 - MAP3K5**
27	Cluster 1	MOS - NF1
	Cluster 2	TRAF6 - MAP3K5
	Cluster 3	**MAPT - DUSP2**
26	Cluster 1	MOS - NF1
	Cluster 2	TRAF6 - MAP3K5
	Cluster 3	MAPT - DUSP2
	Cluster 4	**RELA - MAP2K1**
25	Cluster 1	MOS - NF1
	Cluster 2	TRAF6 - MAP3K5
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1
	Cluster 5	**MAPK3 - CDC25B**
24	Cluster 1	MOS - NF1
	Cluster 2	TRAF6 - MAP3K5

	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1
	Cluster 5	MAPK3 - CDC25B
	Cluster 6	**IL1A – BDNF**
23	Cluster 1	MOS - NF1
	Cluster 2	TRAF6 - MAP3K5 - **PAK2**
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1
	Cluster 5	MAPK3 - CDC25B
	Cluster 6	IL1A – BDNF
22	Cluster 1	MOS - NF1 - **NTRK2**
	Cluster 2	TRAF6 - MAP3K5 - PAK2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1
	Cluster 5	MAPK3 - CDC25B
	Cluster 6	IL1A – BDNF
21	Cluster 1	MOS - NF1 - NTRK2
	Cluster 2	TRAF6 - MAP3K5 - PAK2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1
	Cluster 5	MAPK3 - CDC25B
	Cluster 6	IL1A – BDNF
	Cluster 7	**FGFR4 - MAP3K3**
20	Cluster 1	MOS - NF1 - NTRK2
	Cluster 2	TRAF6 - MAP3K5 - PAK2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1
	Cluster 5	MAPK3 - CDC25B
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - **MKNK2**
19	Cluster 1	MOS - NF1 - NTRK2
	Cluster 2	TRAF6 - MAP3K5 - PAK2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1
	Cluster 5	MAPK3 - CDC25B
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - MKNK2
	Cluster 8	**MAPK11 – PDGFA**
18	Cluster 1	MOS - NF1 - NTRK2
	Cluster 2	TRAF6 - MAP3K5 - PAK2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1
	Cluster 5	MAPK3 - CDC25B
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - MKNK2
	Cluster 8	MAPK11 – PDGFA
	Cluster 9	**MAP3K13 - FGF7**
17	Cluster 1	MOS - NF1 - NTRK2 - **TGFB2**
	Cluster 2	TRAF6 - MAP3K5 - PAK2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1
	Cluster 5	MAPK3 - CDC25B
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - MKNK2
	Cluster 8	MAPK11 – PDGFA

	Cluster 9	MAP3K13 - FGF7
16	Cluster 1	MOS - NF1 - NTRK2 - TGFB2
	Cluster 2	TRAF6 - MAP3K5 - PAK2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1 - **MAP2K2**
	Cluster 5	MAPK3 - CDC25B
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - MKNK2
	Cluster 8	MAPK11 – PDGFA
	Cluster 9	MAP3K13 - FGF7
15	Cluster 1	MOS - NF1 - NTRK2 - TGFB2
	Cluster 2	TRAF6 - MAP3K5 - PAK2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1 - MAP2K2
	Cluster 5	MAPK3 - CDC25B - **RPS6KA1**
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - MKNK2
	Cluster 8	MAPK11 – PDGFA
	Cluster 9	MAP3K13 - FGF7
14	Cluster 1	MOS - NF1 - NTRK2 - TGFB2
	Cluster 2	TRAF6 - MAP3K5 - PAK2 - **FGFR2**
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1 - MAP2K2
	Cluster 5	MAPK3 - CDC25B - RPS6KA1
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - MKNK2
	Cluster 8	MAPK11 – PDGFA
	Cluster 9	MAP3K13 - FGF7
13	Cluster 1	MOS - NF1 - NTRK2 - TGFB2
	Cluster 2	TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1 - MAP2K2
	Cluster 5	MAPK3 - CDC25B - RPS6KA1
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - **DUSP1**
	Cluster 8	MAPK11 – PDGFA - **MKNK2 (changing cluster)**
	Cluster 9	MAP3K13 - FGF7
12	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 – **PRKCB**
	Cluster 2	TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1 - MAP2K2
	Cluster 5	MAPK3 - CDC25B - RPS6KA1
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - DUSP1
	Cluster 8	MAPK11 – PDGFA - MKNK2
	Cluster 9	MAP3K13 - FGF7
11	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - **MAP3K13 (changing cluster)**
	Cluster 2	TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1 - MAP2K2
	Cluster 5	MAPK3 - CDC25B - RPS6KA1
	Cluster 6	IL1A – BDNF
	Cluster 7	FGFR4 - MAP3K3 - DUSP1
	Cluster 8	MAPK11 – PDGFA - MKNK2

	Cluster 9	FGF7
10	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - MAP3K13 - **IL1A - BDNF (changing cluster)- FGF7 (changing cluster)**
	Cluster 2	TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 3	MAPT - DUSP2
	Cluster 4	RELA - MAP2K1 - MAP2K2
	Cluster 5	MAPK3 - CDC25B - RPS6KA1
	Cluster 6	FGFR4 - MAP3K3 - DUSP1
	Cluster 7	MAPK11 – PDGFA - MKNK2
9	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - MAP3K13 - IL1A - BDNF - FGF7
	Cluster 2	TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 3	MAPT - DUSP2 - **FGFR4 - MAP3K3 - DUSP1 (changing cluster)**
	Cluster 4	RELA - MAP2K1 - MAP2K2
	Cluster 5	MAPK3 - CDC25B - RPS6KA1
	Cluster 6	MAPK11 – PDGFA - MKNK2
8	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - MAP3K13 - IL1A - BDNF - FGF7
	Cluster 2	TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 3	MAPT - DUSP2 - FGFR4 - MAP3K3 - DUSP1 - **MKNK2 (changing cluster)**
	Cluster 4	RELA - MAP2K1 - MAP2K2 - **MAPK11 - PDGFA (changing cluster)**
	Cluster 5	MAPK3 - CDC25B - RPS6KA1
7	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - MAP3K13 - IL1A - BDNF - FGF7
	Cluster 2	TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 3	MAPT - DUSP2 - FGFR4 - MAP3K3 - DUSP1 - MKNK2
	Cluster 4	RELA - MAP2K1 - MAP2K2 - MAPK11 - PDGFA
	Cluster 5	MAPK3 - CDC25B - RPS6KA1 - **NFKB1**
6	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - MAP3K13 - IL1A - BDNF - FGF7 - **TRAF6 - MAP3K5 - PAK2 - FGFR2 (changing cluster)**
	Cluster 2	MAPT - DUSP2 - FGFR4 - MAP3K3 - DUSP1 - MKNK2
	Cluster 3	RELA - MAP2K1 - MAP2K2 - MAPK11 - PDGFA
	Cluster 4	MAPK3 - CDC25B - RPS6KA1 - NFKB1
5	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - MAP3K13 - IL1A - BDNF - FGF7 - TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 2	MAPT - DUSP2 - FGFR4 - MAP3K3 - DUSP1 - MKNK2
	Cluster 3	RELA - MAP2K1 - MAP2K2 - MAPK11 - PDGFA - **MAPK9**
	Cluster 4	MAPK3 - CDC25B - RPS6KA1 - NFKB1
4	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - MAP3K13 - IL1A - BDNF - FGF7 - TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 2	MAPT - DUSP2 - FGFR4 - MAP3K3 - DUSP1 - MKNK2 - **RELA - MAP2K1 - MAP2K2 - MAPK11 - PDGFA - MAPK9 (changing cluster)**
	Cluster 3	MAPK3 - CDC25B - RPS6KA1 - NFKB1
3	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - MAP3K13 - IL1A - BDNF - FGF7 - TRAF6 - MAP3K5 - PAK2 - FGFR2

	Cluster 2	MAPT - DUSP2 - FGFR4 - MAP3K3 - DUSP1 - MKNK2 - RELA - MAP2K1 - MAP2K2 - MAPK11 - PDGFA - MAPK9
	Cluster 3	MAPK3 - CDC25B - RPS6KA1 - NFKB1 - **IL1R1**
2	Cluster 1	MOS - NF1 - NTRK2 - TGFB2 - PRKCB - MAP3K13 - IL1A - BDNF - FGF7 - TRAF6 - MAP3K5 - PAK2 - FGFR2
	Cluster 2	MAPT - DUSP2 - FGFR4 - MAP3K3 - DUSP1 - MKNK2 - RELA - MAP2K1 - MAP2K2 - MAPK11 - PDGFA - MAPK9 - **MAPK3 - CDC25B - RPS6KA1 - NFKB1 - IL1R1 (changing cluster)**

Now we define the associations of each gene from Table 2 according to genes that are frequently associated together in the largest number of clusters.

First group: MOS - NF1 - NTRK2 - TGFB2 – PRKCB

The association between MOS and NF1 appeared at the beginning when we applied 29 clusters until NTRK2 joined to them when the number of clusters became 22. In the cluster number 17, TGFB2 joined. The last gene joined to the group was PRKCB at the cluster number 12.

Second group: TRAF6 - MAP3K5 - PAK2 - FGFR2

TRAF6 and MAP3K5 associate together in cluster number 28. Then PAK2 joined when clustering number 23. The last gene joined to the group was FGFR2 at the cluster number 12. This group is repeated until cluster number 6.

Third group: MAPT - DUSP2

MAPT and DUSP2 associated together from cluster number 27 until 10. Then, other genes joined to them.

Fourth group: RELA - MAP2K1 - MAP2K2 - MAPK11 - PDGFA - MAPK9

The association between RELA and MAP2K1 appeared at cluster number 26. Then, MAP2K2 joined them in cluster number 16. This group is repeated until cluster number 8, then MAPK11 and PDGFA joined to them. The last gene joined to the group was MAPK9 at cluster number 5.

Fifth group: MAPK3 - CDC25B - RPS6KA1 - NFKB1 - IL1R1

MAPK3 and CDC25B associate together in cluster number 25. Then RPS6KA1 joined when clustering number 15. In the cluster number 7, NFKB1 is joined. The last gene joined to the group was IL1R1at the cluster number 3.

Sixth group: IL1A – BDNF

IL1A and BDNF associated together from cluster number 24 until 11. Then, they joined to other genes.

Seventh group: FGFR4 - MAP3K3 - DUSP1 - MKNK2

The association between FGFR4 and MAP3K3 appeared at cluster number 21. Then, MKNK2 joined to them in cluster number 20, but in cluster number 13 MKNK2 change his group until cluster number 9 then it returned to this group. MKNK2 belongs to this group because the number of times it appears in this group is more than the other group (it appeared in this group 11 times while in another group just 4 times). The last gene joined to the group was DUSP1 at the cluster number 13.

Eighth group: MAP3K13 - FGF7

The association between MAP3K13 and FGF7 appeared at cluster number 18 until 12. Then, they joined to other genes.

B. Linear Regression Experiment

After we extracted the gene groups, we applied the linear regression method to interpret the correlation coefficients. In WEKA, we adjusted some settings to know the correlation coefficients among all genes in one group. When we selected the linear regression method, we set the attribute selection method field to NO attribute. Table 2 shows the results:

Table 2. A Result Of Experiment Using Linear Regression

Group Number	Y	Xs	Correlation Value (β)
1	**MOS**	PRKCB	-0.25
		NF1	+0.5545
		TGFB2	-0.0233
		NTRK2	+0.2113
	PRKCB	MOS	-0.1207
		NF1	+0.8328
		TGFB2	+0.0741
		NTRK2	+0.1635
	NF1	MOS	+0.2351
		PRKCB	+0.7315
		TGFB2	+0.1283
		NTRK2	-0.0953
	TGFB2	MOS	-0.0511
		PRKCB	+0.3363
		NF1	+0.6625
		NTRK2	+0.3698
	NTRK2	MOS	+0.3622
		PRKCB	+0.5807
		NF1	-0.3851
		TGFB2	+0.2895
2	**TRAF6**	PAK2	-0.0033
		FGFR2	+0.2302
		MAP3K5	+0.1608
	PAK2	TRAF6	-0.035
		FGFR2	-0.3851
		MAP3K5	-0.1115
	FGFR2	TRAF6	+0.7766
		PAK2	-0.1234
		MAP3K5	-0.3446
	MAP3K5	TRAF6	+0.90
		PAK2	-0.168
		FGFR2	-0.90
3	**DUSP2**	MAPT	+0.2872
	MAPT	DUSP2	+0.7159
4	**MAPK11**	RELA	+0.0624
		PDGFA	+0.5131
		MAP2K1	-0.1363
		MAP2K2	-0.0606
		MAPK9	-0.1693
	RELA	MAPK11	+0.175
		PDGFA	+0.0272
		MAP2K1	-0.0059
		MAP2K2	+0.3966
		MAPK9	-0.0204
	PDGFA	MAPK11	+0.9894
		RELA	+0.0187
		MAP2K1	-0.2666
		MAP2K2	-0.2754
		MAPK9	+0.0528
	MAP2K1	MAPK11	-0.2148
		RELA	-0.0033

		PDGFA	-0.2179
		MAP2K2	-0.322
		MAPK9	-0.1184
	MAP2K2	MAPK11	-0.058
		RELA	+0.1354
		PDGFA	-0.1368
		MAP2K1	-0.1957
		MAPK9	-0.2321
	MAPK9	MAPK11	-0.9496
		RELA	-0.0408
		PDGFA	+0.1536
		MAP2K1	-0.4216
		MAP2K2	-0.90
5	**MAPK3**	RPS6KA1	+0.1622
		CDC25B	+0.1976
		IL1R1	-0.0957
		NFKB1	+0.1103
	RPS6KA1	MAPK3	+0.1561
		CDC25B	+0.118
		IL1R1	+0.0215
		NFKB1	+0.0773
	CDC25B	MAPK3	+0.8836
		RPS6KA1	+0.5479
		IL1R1	+0.0616
		NFKB1	+0.2324
	IL1R1	MAPK3	-0.90
		RPS6KA1	+0.2643
		CDC25B	+0.1634
		NFKB1	-0.0921
	NFKB1	MAPK3	+0.0909
		RPS6KA1	+0.0662
		CDC25B	+0.0428
		IL1R1	-0.0064
6	**IL1A**	BDNF	+0.7658
	BDNF	IL1A	+0.7731
7	**DUSP1**	MKNK2	+0.1864
		FGFR4	+0.1538
		MAP3K3	+0.2045
	MKNK2	DUSP1	+0.5298
		MAP3K3	+0.4731
		FGFR4	+0.2693
	FGFR4	DUSP1	+0.0882
		MKNK2	+0.0544
		MAP3K3	+0.9066
	MAP3K3	DUSP1	+0.0681
		MKNK2	+0.0554
		FGFR4	+0.5262
8	**MAP3K13**	FGF7	+0.7071
	FGF7	MAP3K13	+0.5845

C. Graphical Representation for Genes Associations

After we applied the K-means method, we got the associations between genes in our dataset. Then, in the linear regression method, we specified the amount of those associations. One of the

most representative methods for gene associations is a graphical model representation. Based on our results, we plotted Fig.3 which shows genes association produced from the K-means method. The numerated arrows in this figure represent the amount of those associations produced from the linear regression method.

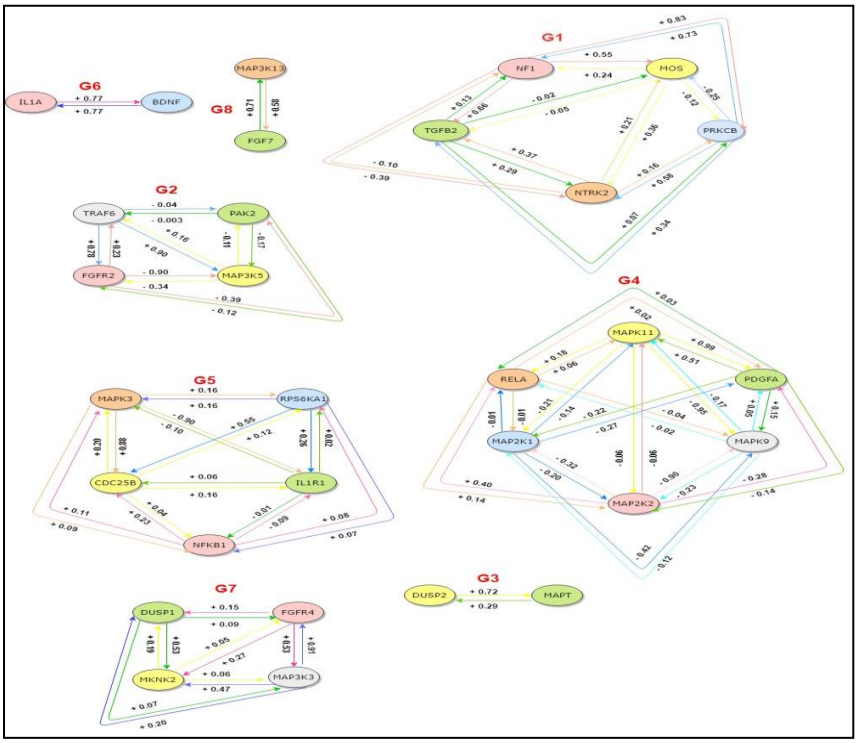

Fig.3 Graphical Representation for Genes Associations

The most significant finding based on the results discussed above is the correlation between two genes, similar in signals while the numbers are different in most cases. We explain those correlations for each group in detail.

G1: MOS and NF1have a positive correlation and the effect of NF1 on MOS more because it has the highest correlation's value. MOS and PRKCB have a negative correlation and a weak effect on some of them because they have a low correlation value. MOS and TGFB2 have the same correlation between MOS and PRKCB. MOS and NTRK2 have a positive correlation and a weak effect on some of them because they have a low correlation value.

NF1 and TGFB2 have a positive correlation and the effect of NF1 on TGFB2 more because it has the highest correlation's value. NF1 and NTRK2 have a negative correlation and a weak effect on some of them because they have a low correlation value. NF1 and PRKCB have a positive correlation and a strong effect on some of them because they have a high correlation value.

TGFB2 and NTRK2 have a positive correlation and a weak effect on some of them because they have a low correlation value. TGFB2 and PRKCB have the same correlation between TGFB2 and NTRK2.

NTRK2 and PRKCB have a positive correlation and effect of PRKCB on NTRK2 more because it has the highest correlation's value.

G2: TRAF6 and PAK2 have a negative correlation and a weak effect on some of them because they have a low correlation value. TRAF6 and FGFR2 have a positive correlation and the effect of TRAF6 on FGFR2 more because it has the highest correlation's value. TRAF6 and MAP3K5 have a positive correlation and the effect of TRAF6 on MAP3K5 more because it has the highest correlation's value.

PAK2 and MAP3K5 have a negative correlation and a weak effect on some of them because they have a low correlation value. PAK2 and FGFR2 have the same correlation between PAK2 and MAP3K5.

MAP3K5 and FGFR2 have a negative correlation and the effect of FGFR2 on MAP3K5 more because it has the highest correlation's value.

G3: MAPT and DUSP2 have a positive correlation and the effect of DUSP2 on MAPT more because it has the highest correlation's value.

G4: MAPK11 and RELA have a positive correlation and a weak effect on some of them because they have a low correlation value. MAPK11 and PDGFA have a positive correlation and the effect of MAPK11 on PDGFA more because it has the highest correlation's value. MAPK11 and MAP2K1have a negative correlation and a weak effect on some of them because they have a low correlation value. MAPK11 and MAP2K2 have a negative correlation and have the same effect on some of them, this effect is weak because they have a low correlation value. MAPK11 and MAPK9 have a negative correlation and the effect of MAPK11 on MAPK9 more because it has the highest correlation's value.

RELA and MAP2K1 have a negative correlation and have the same effect on some of them, this effect is weak because they have a low correlation. RELA and MAP2K2 have a positive correlation and a weak effect on some of them because they have a low correlation value. RELA and PDGFA have the same correlation between RELA and MAP2K2. RELA and MAPK9 have a negative correlation and a weak effect on some of them because they have a low correlation value.

MAP2K1 and MAP2K2 have a negative correlation and a weak effect on some of them because they have a low correlation value. MAP2K1 and MAPK9 have the same correlation between MAP2K1 and MAP2K2. MAP2K1 and PDGFA also have the same correlation between MAP2K1 and MAP2K2.

MAP2K2 and MAPK9 have a negative correlation and the effect of MAP2K2 on MAPK9 more because it has the highest correlation's value. MAP2K2 and PDGFA have a negative correlation and a weak effect on some of them because they have a low correlation value.

MAPK9 and PDGFA have a positive correlation and a weak effect on some of them because they have a low correlation value.

G5: MAPK3 and RPS6KA1 have a positive correlation and have the same effect on some of them, this effect is weak because they have a low correlation. MAPK3 and CDC25B have a positive correlation and the effect of MAPK3 on CDC25B more because it has the highest correlation's value. MAPK3 and IL1R1 have a negative correlation and the effect of MAPK3 on IL1R1 more because it has the highest correlation's value. MAPK3 and NFKB1 have a positive correlation and a weak effect on some of them because they have a low correlation value.

RPS6KA1 and IL1R1 have a positive correlation and a weak effect on some of them because they have a low correlation value. RPS6KA1 and CDC25B have a positive correlation and the effect of RPS6KA1 on CDC25B more because it has the highest correlation's value. RPS6KA1 and NFKB1 have a positive correlation and a weak effect on some of them because they have a low correlation value. IL1R1 and CDC25B have a positive correlation and a weak effect on some of them because they have a low correlation value. IL1R1 and NFKB1 have a negative correlation and a weak effect on some of them because they have a low correlation value.

CDC25B and NFKB1 have a positive correlation and a weak effect on some of them because they have a low correlation value.

G6: IL1A and BDNF have a positive correlation and have the same effect on some of them, this effect is strong because they have a high correlation value.

G7: DUSP1 and FGFR4 have a positive correlation and a weak effect on some of them because they have a low correlation value. DUSP1 and MAPK3K3 have the same correlation between DUSP1 and FGFR4. DUSP1 and MKNK2 have a positive correlation and the effect of DUSP1 on MKNK2 more because it has the highest correlation's value.

FGFR4 and MAPK3K3 have a positive correlation and the effect of MAPK3K3 on FGFR4 more because it has the highest correlation's value. FGFR4 and MKNK2 have a positive correlation and a weak effect on some of them because they have a low correlation value.

MAPK3K3 and MKNK2 have a positive correlation and the effect of MAPK3K3 on MKNK2 more because it has the highest correlation's value.

G8: MAPK3K13 and FGF7 have a positive correlation and a strong effect on some of them because they have a high correlation value.

5. CONCLUSION

Breast Cancer is a complex disease and early detection is essential for effective treatment. In our paper we present the gene expression correlations in graphical model using the K-means clustering and linear regression methods. The K-means clustering is used to find out the best subset of genes then the linear regression method is applied to predict the amount of association between those genes. With this approach, doctors can apply the target cancer treatment to the infected area instead of applying the treatment to the whole body, which normally results in effecting all cells even normal once. The work we have done in this project was based on 30 cancer samples. Our future goal is work with the biggest dataset. Another possible direction for future work is concerned with a comparison between cancer and normal samples to find out which gene works in a different way. For example, based on our results, MAPK3K13 and FGF7 have a positive correlation and a strong effect on some of them. We want to know how MAPK3K13 and FGF7 work in normal samples.

REFERENCES

[1] V. Chaurasia and S. Pal, "A novel approach for breast cancer detection using data mining techniques," 2017.

[2] M. A. Hasan, M. K. Hasan, and M. A. Mottalib, "Linear regression–based feature selection for microarray data classification," International journal of data mining and bioinformatics, vol. 11, no. 2, pp. 167-179, 2015.

[3] W. P. Kuo, E.-Y. Kim, J. Trimarchi, T.-K. Jenssen, S. A. Vinterbo, and L. Ohno-Machado, "A primer on gene expression and microarrays for machine learning researchers," Journal of Biomedical Informatics, vol. 37, no. 4, pp. 293-303, 2004.

[4] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Computer Science, vol. 83, pp. 1064-1069, 2016.

[5] M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, "Machine learning models in breast cancer survival prediction," Technology and Health Care, vol. 24, no. 1, pp. 31-42, 2016.

[6] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and structural biotechnology journal, vol. 13, pp. 8-17, 2015.

[7] Y. Zhou, R. Qureshi, and A. Sacan, "Data simulation and regulatory network reconstruction from time-series microarray data using stepwise multiple linear regression," Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 1, no. 1-2, pp. 3-17, 2012.

[8] P. A. v. D. S.J. Van Laere, L.Y. Dirix, P.B. Vermeulen "Microarray data analysis: strategies, pitfalls and applications in clinical oncology," 2011.

[9] Q. Sheng, Y. Moreau, F. De Smet, K. Marchal, and B. De Moor, "Cluster analysis of microarray data," ed: unpublished.

[10] T. D. Pham, C. Wells, and D. I. Crane, "Analysis of microarray gene expression data," Current Bioinformatics, vol. 1, no. 1, pp. 37-53, 2006.

[11] A. Schulze and J. Downward, "Navigating gene expression using microarrays—a technology review," Nature cell biology, vol. 3, no. 8, p. E190, 2001.

[12] J. Quackenbush, "Microarray data normalization and transformation," Nature genetics, vol. 32, p. 496, 2002.

[13] S. Turgut, M. Dağtekin, and T. Ensari, "Microarray breast cancer data classification using machine learning methods," in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018, pp. 1-3: IEEE.

[14] W. Dubitzky, M. Granzow, and D. Berrar, "Data mining and machine learning methods for microarray analysis," in Methods of microarray data analysis: Springer, 2002, pp. 5-22.

[15] M. Pirooznia, J. Y. Yang, M. Q. Yang, and Y. Deng, "A comparative study of different machine learning methods on microarray gene expression data," BMC genomics, vol. 9, no. 1, p. S13, 2008.

[16] S.-B. Cho and H.-H. Won, "Machine learning in DNA microarray analysis for cancer classification," in Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19, 2003, pp. 189-198: Australian Computer Society, Inc.

[17] R. Jin, L. Si, S. Srivastava, Z. Li, and C. Chan, "A knowledge driven regression model for gene expression and microarray analysis," in Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, 2006, pp. 5326-5329: IEEE.

[18] R. Radha and P. Rajendiran, "Using K-Means clustering technique to study of breast cancer," in Computing and Communication Technologies (WCCCT), 2014 World Congress on, 2014, pp. 211-214: IEEE.

[19] S. Datta and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data," Bioinformatics, vol. 19, no. 4, pp. 459-466, 2003.

[20] M. Zhao, Y. Tang, H. Kim, and K. Hasegawa, "Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer," Cancer informatics, vol. 17, p. 1176935118810215, 2018.

[21] R. Suresh, K. Dinakaran, and P. Valarmathie, "Model based modified k-means clustering for microarray data," in 2009 International Conference on Information Management and Engineering, 2009, pp. 271-273: IEEE.

[22] M. West et al., "Predicting the clinical status of human breast cancer by using gene expression profiles," Proceedings of the National Academy of Sciences, vol. 98, no. 20, pp. 11462-11467, 2001.

[23] L. X. Qin and S. G. Self, "The clustering of regression models method with applications in gene expression data," Biometrics, vol. 62, no. 2, pp. 526-533, 2006.

[24] L.-X. Qin, L. Breeden, and S. G. Self, "Finding gene clusters for a replicated time course study," BMC research notes, vol. 7, no. 1, p. 60, 2014.

[25] R. Sharma, M. A. Alam, and A. Rani, "K-means clustering in spatial data mining using weka interface," International Journal of Computer Applications, pp. 26-30, 2012.

[26] D. C. Montgomery and G. C. Runger, Applied statistics and probability for engineers. John Wiley & Sons, 2010.

[27] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[28] S. Singhal and M. Jena, "A study on WEKA tool for data preprocessing, classification and clustering," International Journal of Innovative technology and exploring engineering (IJItee), vol. 2, no. 6, pp. 250-253, 2013.

[29] J. C. Chang et al., "Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer," The Lancet, vol. 362, no. 9381, pp. 362-369, 2003.