Prediction of the Thromboembolic Syndrome: an Application of Artificial Neural Networks in Gene Expression Data Analysis

Mahdieh Khalili [1] , Hamid Alavi Majd [1,*] , Soheila Khodakarim [1], Batool Ahadi [1] Mohsen Hamidpour [2]

[1] Department of Biostatistics, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Tehran, Iran
[2] Department of Hematology, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Tehran, Iran

* Corresponding Author: email address: alavimajd@gmail.com (H. Alavi Majd)

ABSTRACT

The aim of this study was to propose a method for improving the power of recognition and classification of thromboembolic syndrome based on the analysis of gene expression data using artificial neural networks. The studied method was performed on a dataset which contained data about 117 patients admitted to a hospital in Durham in 2009. Of all the studied patients, 66 patients were suffering from thromboembolic syndrome and 51 people were enrolled in the study as the control group. The gene expression level of 22277 was measured for all the samples and was entered into the model as the main variable. Due to the high number of variables, principal components analysis and auto-encoder neural network methods were used in order to reduce the dimension of data. The results showed that when using auto-encoder networks, the classification accuracy was 93.12. When using the PCA method to reduce the size of the data, the obtained accuracy was 78.26, and hence a significant difference in the accuracy of classification was observed. If auto-encoder network method is used, the sensitivity and specificity will be 92.58 and 93.68 and when PCA method is used, they will be 0.77 and 0.78 respectively. The results suggested that auto-encoder networks, compared with the PCA method, had a higher level of accuracy for the classification of thromboembolic syndrome status.

Keywords: Thromboembolic syndrome; gene expression data; principal component analysis (PCA);auto-encoder neural networks.

INTRODUCTION

Venous thromboembolism (VTE)syndrome is a disease which includes deep vein thrombosis (DVT) and pulmonary embolism (PE). Embolism is the penetration of something into the bloodstream and its subsequent movement in the direction of blood circulation; in addition, thromboembolism is an embolism caused by a blood clot. When a blood clot is formed within a blood vessel, its movement in the blood and its transfer to another place is called thromboembolism[1]. Moreover, deep vein thrombosis occurs when the blood in deep veins moves more slowly than usual or when there are factors that increase the tendency of blood to form a clot; furthermore, when the internal layer of the vein is damaged, the probability of deep vein thrombosis also increases. The formation of clots in the inner lining of deep veins is dangerous, because these clots may break off and enter the bloodstream and they may block the important arteries, more specificallythe major arteries of the lungs and lead to permanent damage or death [2].

This disorder, which is one of the most important threats to life, occurs in half of the patients either admitted or not admitted to hospitals. More than 30% of cases of thromboembolic syndromes can relapse, and neglecting this important fact can cause long-term complications such as high blood pressure, chronic thromboembolic pulmonary hypertension (CTPH) and post-thrombotic syndrome (PTS). Approximately, 70,000 cases of venous thromboembolism are hospitalized in the UK, of whom 12% die during hospitalization period and about 30% die after three years [3].

All the predisposing factors of thrombosis have not been known yet and the available tools are not able to identify more than half of the predisposing factors of thromboembolism. On the other hand, the presence of pre-coagulation or thrombophilia factors is not necessarily a sign for the occurrence of thrombosis. In most clinical thrombosis syndromes, the presence of one or more inherited condition (lack of protein c) together with one or more acquired factors (pregnancy, immobilization, surgery) can move the hemostasis system toward thrombosis. After the elimination or treatment of the acquired factor, due to the presence of inherited thrombophilia factor, the homeostasis system will again go to a subclinical state. The asymptomatic state may remain the same for the rest of life, or may clinically manifest itself again after the development of a new underlying acquired disease. Unfortunately, due to inflammatory, destructive and reconstructive changes in the damaged vascular endothelium, every case of thrombosis is considered as an acquired factor involved in the recurrence of thrombosis [4].

On the one hand, since many years ago the researches in the medical field have shown that some mutations can cause cancers and various diseases. As a consequence, studies on gene expression are of great importance. DNA microarray technology provides the possibility of studies on genes. In recent years, this technology has had a key role in biomedical research and has led to major developments in biological and biomedical fields. Hence, it is a field of interest to many researchers. DNA microarray is a collection of microscopic DNA spots that are connected to a solid surface such as glass, plastic, or silicon chip and form an array. Researchers have successfully used microarray techniques to measure the expression levels of large numbers of genes simultaneously in a variety of genomics analyses including drug discovery, genes identification, and clinical diagnosis [5].

Three general categories of studies which are conducted based on microarray technology are: class comparison, class discovery, and class prediction studies. Class comparison studies are intended to compare the gene expression profiles of two or more groups of patients. This type of study is aimed to identify the genes that have different expression levels in the two groups [6]. In class discovery studies, the aim is to discover subgroups that have a similar gene expression profiles in a dataset [7]. In class prediction studies, the aim is to investigate pre-defined categories, for example, certain types of patients with cancer and healthy individuals. The aim of this study is to use the gene expression profiles and distinguish the category of each case [8]. To perform such analyses we use statistical methods such as discriminate analysis and machine learning [9].

In this study, we used artificial neural networks as a data mining algorithm and conducted a class prediction study. In other words, we used artificial neural network models and examined whether it is possible to identify patients with syndrome and distinguish them from healthy people using gene expression data of the samples.

One of the most important challenges of the microarray data analysis is the imbalance between the number of variables (the level of gene expression) and the number of samples available. In the study of diseases, for instance the case of special diseases, because of problems with sample collection and testing cost, the number of samples available is very limited which leads to an imbalance between the number of variables and the number of cases[10]. As a result, when using the standard methods of machine learning, we may face the problem of the low number of samples. In other words, the gene expression matrix will have a very high volume of genes and a very limited number of samples, and this results in imbalance between the number of rows and columns of the matrix; thus, in turn, it will lead to very high computational complexity and reduced functionality of classification tools. This problem which is caused due to the high volume of features and low number of samples is known as the small sample size or SSS [11].

To overcome this problem, various methods of dimension reduction can be used, among which feature selection methods and feature extraction techniques are the note-worthy ones. Feature selection methods only select a few features as the superior features and ignore the other ones. Feature extraction methods use linear or non-linear combination of all the features available and produce smaller sets of features [12].

As one of the disadvantages of feature selection methods, they ignore a number of features that may contain valuable information about their classification. In addition, usually the selection of a number of features is optional and from a theoretical point of view, it is not easy to select a specific number of features to reach an optimal performance. The desired number of features is typically determined through empirical methods [11]. For the classification purposes, feature extraction methods have a better performance than feature selection methods, because obtained features are in fact a linear or non-linear combination of basic features and can cover a large part of the variance in the original data.

Principle component analysis (PCA) is one of the most common methods of feature extraction. The use of PCA method helps to reduce the dimensions of variables without losing the data variables have in their covariance matrix [13]. Gene expression data are a type of data that contain a lot of variables which are strongly correlated with one another. Therefore, the use of this method helps to reduce the number of variables (genes); yet, this method only includes linear transformations of data while for the majority of the datasets, especially for gene expression data, we need a method that discovers non-linear relationships too.

Auto-encoder neural networks, first proposed by Hinton in 1980, are a type of artificial neural networks that can take inputs to the network and produce an output with the least amount of deviation. In fact, by entering the inputs to the network, they can compress or encode the data as much as possible without loss of data and hence they can generate a shorter presentation of data. Then they try to reconstruct or recode the compressed data. The goal of training the network is to reduce the rate of error in data recovery, so as to achieve the most efficient compressed set of primary data [14]. Thus, in this study, we tried to examine the status of susceptibility to the disease through the selection of variables or the appropriate genes by reducing the size of data and using one of the classifiers.

MATERIALS AND METHODS

In this applied study, the data from Luis et al.'s (2011) study was used. The data are related to the gene expression of 66 patients with thromboembolism and 51 healthy individuals as the control group who had referred to a hospital in the city of Durham in 2009. All the information and datasets are available to the public at http://www.ncbi.nlm.nih.gov/geo. The data have been normalized after being downloaded from the GEO website and sent to babelomics website, using the RMA algorithm [15]. The normalized data were inserted into a 117 × 22277 gene expression matrix, and the analyses and the fitness of the model were completed using MATLAB software.

The approach explored in this study consists of two phases. The first phase, also called "feature learning," includes the dimension reduction of the variables, and second phase is known as "classification" or "classifier learning" phase.

The first phase, also called dimension reduction phase, involves two stages. The first stage uses PCA and adds a number of random variables. The second stage includes entering data obtained from the PCA into an auto-encoder neural network. Thus, to reduce the dimensions of the variables, Karhunen–Loève Transform (K-L) was first used, reducing the number of variables from 22277 to 116. This transform is, in fact, a technique based on PCA which, by applying some transformations in the feature values of PCA, can resolve the problem where the number of variables is a hundred times that of samples [16].

The higher number of variables with respect to that of samples may cause problems in the computations related to the feature vector. Therefore, after applying this transform, the dimensions of gene expression matrix would be reduced to 117 × 116. Then, 200 variables from among the initial variables would be selected randomly and added to the 116 components obtained from K-L transform. The purpose of adding these variables is to increase the chance of detecting and discovering the existing non-linear relationships that are latent in the variables obtained from PCA.

The outcome would, thus, be a 117 × 316 matrix ready for entering into an auto-encoder network. At the second stage, the resulting 316 variables would be entered into an auto-encoder network in order to further reduce the dimensions. The auto-encoder network would compress the variables. These variables adequately represent the initial 22277 gene expression variables and could be used for classification in the second phase of the study.

In the second phase, the classification was done using a Radial Basis Function (RBF) network, and the accuracy, sensitivity, and specificity were also calculated. In this phase, two different methods were employed for selecting the cluster centers: Greedy Search and K-means. Furthermore, the accuracy of the classification was also measured in order to examine and compare the performance of auto-encoder networks with the cases in which only K-L transform has been applied for reducing the dimensions.

This study includes only one independent variable, that is, the amount of gene expression; this is a continuous, quantitative value and is calculated using microarray technology. The response variable is a binary, qualitative variable in which 1 represents the presence of the syndrome, and 0 indicates its absence in the sample members.

The structure of the auto-encoder network applied in the study is shown in figure 1.

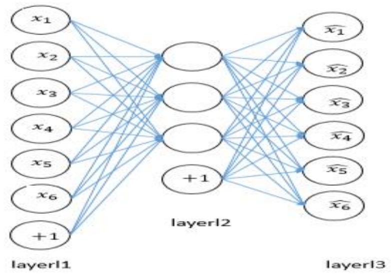

Figure 1.Structure of the auto-encoder network

Like all types of neural networks, auto-encoder networks also aim at minimizing cost function value. Cost function, as a fundamental concept in network training, determines the difference between the available and the optimal solutions. In auto-encoder networks, the desired value is achieved when the input values of the network are exactly produced in the output layer. In other words, an auto-encoder attempts to train the function $h_{w,b}(x)$, which is the output value of the network, so we have:

$$h_{w,b}(x) \approx x$$

x is the very input value of the network[17].

RESULTS

As mentioned earlier, the duty of auto-encoder network is to reduce the dimensions of input variables so that the new compressed set could adequately represent the initial variables. However, determining the number of these variables is arbitrary and is achieved by trial and error. In addition, when doing classifications using the RBF neural networks, the selection of the centers of clusters is done with respect to the number of variables entered into the network. Therefore, the accuracy of measurement was calculated after selecting various numbers of input variables for the RBF network and setting1/5, 1/10, and 1/20 ratios for the number of cluster centers. In Table 1, the number of nodes indicates the number of variables entered into the RBF network. For instance, 20 means after the auto-encoder network is used, 20 variables would be selected as the compressed representatives of the variables; these 20 variables would, then, be entered into the RBF network for the purpose of classification. ACC1, ACC2, and ACC3 in the table show, respectively, the accuracy levels of the network classification when the numbers of the centers of the clusters were 1/5, 1/10, and 1/20 times the number of the input variables and when the Greedy Search was used for selecting the centers of the clusters.

Table 1. Accuracy of classification in the Greedy Search method

number of nodes	acc1	acc2	acc3
20	55.76	52.24	74.49
30	49.49	60.65	90.54
40	54.75	79.34	87.17
50	73.48	85.54	90.57
60	69.31	87.10	92.35
70	79.71	84.67	93.15
80	81.99	91.41	93.94
90	89.71	90.68	93.12
100	87.10	92.24	92.35
110	89.78	89.67	94.02

As Figure 2 shows, in cases where the numbers of the centers of the clusters are selected as 1/5, 1/10, and 1/20 times the number of the input variables, as the number of input variables increases, the accuracy of classification also improves.

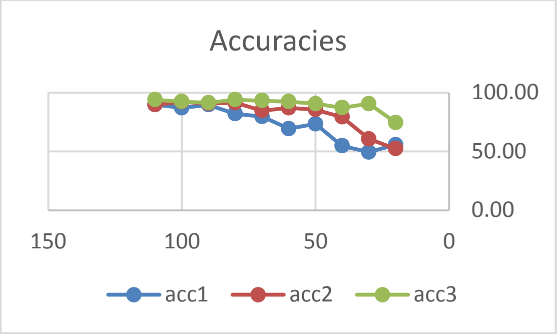

Figure 2. Plot of changes in the accuracy levels of classification in the Greedy Search method

We now calculate the accuracy values using the K-means method. Table 2 reveals the calculated accuracy values obtained using the K-means method. In this method, as also evident in Figure 3, as the number of input variables increases, the accuracy of classification also increases. However, in both methods (viz., Greedy Search and K-means), after 80 features are applied, the accuracy values become stable as no significant changes in the values is observed. Thus, selecting 80, 90, and 100 features produces almost identical accuracy values.

Table 2. Accuracy of classification in the K-means method

number of nodes	acc1	acc2	acc3
20	55.76	53.94	87.21
30	54.71	64.92	91.41
40	53.91	85.39	87.21
50	68.26	87.28	93.15
60	76.78	86.26	87.31
70	85.54	87.97	92.31
80	89.23	88.87	92.28
90	92.28	91.44	93.49
100	91.25	91.41	94.89
110	92.31	88.80	95.76

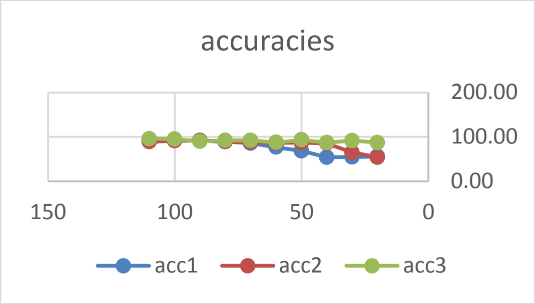

Figure 3. Plot of changes in the accuracy levels of classification in the K-means method

According to the results obtained, all three criteria of 80, 90, and 100 variables could be regarded as appropriate selections adequately representing the initial variables. However, the calculation of the cost function (which is, as stated earlier, a function of the difference between the network input and output variables) indicated that the most optimal results would be achieved when selecting 90 variables.Therefore, it was revealed that using dimension reduction methods, the dimensions of the 22277 genes existing in the study could be reduced to only 90 variables that adequately represent the initial variables. The sensitivity and specificity values of the model are 92.58 and 93.68, respectively. Moreover, to investigate whether using auto-encoder networks would improve the outcome of the classification; the accuracy level is dealt with under the condition that only the K-L method of PCA is used in the dimension reduction phase. That is, after the data are compressed at this stage, auto-encoder networks are no longer used for further compression. In such a case, the sensitivity and specificity values are 0.77 and 0.78, respectively.

The accuracy of prediction, using auto-encoder network is 93.12, and accuracy of prediction, without using auto-encoder network is 78.26; the comparison of the measurement accuracy values in the two models applied indicates that using auto-encoder networks would result in a significant difference in the accuracy level.

To investigate the generalizability of the model with the other datasets, the other 12 gene expression datasets were used (Table 3), and the prediction accuracy level was calculated. From among the 12 datasets that were used in this study, the proposed model showed a better performance in ten cases when compared with the basic method (PCA)

Table 3. Characteristics of gene expression datasets

No.	Dataset	Number of Variables or Genes	Number of Samples
1	Thromboembolism	22277	117
2	AML (Mills et al., 2009)	54613	183
3	Adenocarcinoma (Fujiwara et al.,2011)	34749	28
4	Breast Cancer (Woodward et al.,2013)	30006	20
5	Leukemia (Cheok et al., 2003)	12600	60
6	AML (Yagi et al., 2003)	12625	27
7	Seminoma (Gashaw et al., 2005)	12625	20
8	Ovarian Cancer (Petricoin et al.,2002)	15154	153
9	Colon Cancer (Alon et al., 1999)	2000	62
10	Medulloblastoma (Pomeroy et al.,2002)	7129	30
11	Prostate Cancer (Singh et al., 2002)	12600	34
12	Leukemia (Verhaak et al., 2009)	54613	230

The calculated accuracy values related to these datasets are presented in Table 4.

Table 4. Accuracy values of the model using an auto-encoder network and the basic method (PCA)

No.	Ratio of the Number of Samples to that of Features × 100	With Auto-Encoder	Without Auto-Encoder
1	0.525	93.12	78.26
2	0.335	74.37	72.86
3	0.0805	91.67	80.48
4	0.066	56.09	40.27
5	0.476	75.68	70.22
6	0.213	81.67	75.62
7	0.158	35.28	52.25
8	1.009	77.14	89.15
9	3.1	87.05	91.28
10	0.420	66.67	64.95
11	0.269	77.48	64.75
12	0.421	89.76	82.16

DISCUSSION

On the whole, from among the 12 datasets that were used in this study, the proposed model performed better in 9 cases when compared with the basic method (PCA). As mentioned earlier, in three of all datasets, the basic method showed a better performance in comparison with the model proposed in the study. Further examinations of the data revealed that, in two cases of these datasets Alon et al.'s study of colorectal cancer[18] and Petricoin et al.'s study of ovarian cancer[19], when compared with other datasets, the number of samples and the number of features are relatively closer to one another. For this reason, the ratio of the number of samples to the number of features× 100 was calculated for each of the 12 datasets. The ratios for the ovarian cancer and colorectal cancer datasets are 1.010 and 3.100, respectively; these are higher ratios compared with those for other datasets. Thus, the better performance of the basic model, in comparison with the proposed model (i.e., the auto-encoder network), could be attributed to the higher ratio of the number of samples to that of features compared with other studies.

Furthermore, the examinations show that when considering the datasets with lower ratios. for example theFujiwara et al.'s study of adenocarcinoma[20] with the sample to feature ratio of 0.081 or Singh et al.'s study of the prostate cancer[21] with the ratio of 0.26, auto-encoder networks produce significantly improved results when compared with PCA alone.

For the seminoma dataset in the Gashaw et.al's study[22], although the calculated accuracy value for the basic model was higher than that of the model proposed in the study, it is not related to the higher ratio of the number of samples to the number of features (as it was the case of the ovarian cancer and colorectal cancer datasets). Rather, as stated before, none of the models performed well. In Fakoor et al.'s study[14] on this dataset using auto-encoders networks along with classification by support vector machines, the model also performed poorly in the final classification so that using various types of auto-encoder networks did not result in an accuracy level of more than 56% in that study.

In the end, considering the fact that not all the causal factors of thrombosis are known yet and that the available tools are simply unable to identify half of the causal factors of thromboembolism, it could be stated that using peoples' gene expressions while applying the proposed model could help diagnose people suffering from the syndrome with a higher level of accuracy.

The potential problems with most feature selection methods are scalability and generality of features. For example, Aliferis et al. used recursive feature elimination and univariate association filtering approaches to select a small subset of the gene expressions as a reduced feature set[23] or Ramaswamy et al. applied recursive feature elimination using SVM to find similarly a small number of gene expressions to be used as the feature space for the classification[24]. In these methods there is no possibility of applying data from various types of cancer to automatically form features which help to enhance the detection and diagnosis of a specific one: for example prostate cancer data cannot be used in selecting features for breast cancer detection, reducing the basis for feature learning. In contrast to these methods, our proposed method can use data from different cancer types in the feature learning step, promising the potential for effective feature learning in the presence of very limited data sets.

ACKNOWLEDGEMENT

This paper is based on a proposal from Department of Biostatistics, Faculty of Paramedical Sciences. Thereby, the effort of the chief of faculty is really appreciated for providing required facilities of the present research.

"The authors declare no conflict of interest"

REFRENCES

1.Marik PE, Plante LA. Venous thromboembolic disease and pregnancy. New England Journal of Medicine. 2008;359(19):2025-33.

2.Brunner LS, Smeltzer SCC, Bare BG, Hinkle JL, Cheever KH. Brunner & Suddarth's textbook of medical-surgical nursing: Lippincott Williams & Wilkins; 2010.

3.Qaseem A, Chou R, Humphrey LL, Starkey M, Shekelle P. Venous thromboembolism prophylaxis in hospitalized patients: a clinical practice guideline from the American College of Physicians. Annals of internal medicine. 2011;155(9):625-32.

4.Sharifiyan R, Ravanbod M. Coagulating like conditions–thrombophilia. 2005.

5.Beltrame F, Papadimitropoulos A, Porro I, Scaglione S, Schenone A, Torterolo L, et al. Gemma—a grid environment for microarray management and analysis in bone marrow stem cells experiments. Future Generation Computer Systems. 2007;23(3):382-90.

6.Beer DG, Kardia SL, Huang C-C, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature medicine. 2002;8(8):816-24.

7.Offenbacher S, Lieff S, Boggess K, Murtha A, Madianos P, Champagne C, et al. Maternal periodontitis and prematurity. Part I: Obstetric outcome of prematurity and growth restriction. Annals of periodontology. 2001;6(1):164-74.

8.Hwang K-B, Cho D-Y, Park S-W, Kim S-D, Zhang B-T. Applying machine learning techniques to analysis of gene expression data: cancer diagnosis. Methods of Microarray Data Analysis: Springer; 2002. p. 167-82.

9.Dubitzky W, Granzow M, Berrar D. Comparing symbolic and subsymbolic machine learning approaches to classification of cancer and gene identification. Methods of Microarray Data Analysis: Springer; 2002. p. 151-65.

10.Xu R, Damelin S, Nadler B, Wunsch DC. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps. Artificial intelligence in medicine. 2010;48(2):91-8.

11.Sharma A, Paliwal KK. Cancer classification by gradient LDA technique using microarray gene expression data. Data & Knowledge Engineering. 2008;66(2):338-47.

12.Kittler J. Feature selection and extraction. Handbook of pattern recognition and image processing. 1986:59-83.

13. Johnson RA, Wichern DW. Applied multivariate statistical analysis: Prentice hall Englewood Cliffs, NJ; 1992.

14.Fakoor R, Ladhak F, Nazi A, Huber M, editors. Using deep learning to enhance cancer diagnosis and classification. Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare Atlanta, Georgia: JMLR: W&CP; 2013.

15. Alavi-Majd H, Khodakarim S, Zayeri F, Rezaei-Tavirani M, Tabatabaei SM, Heydarpour-Meymeh M. Assessment of gene

set analysis methods based on microarray data. Gene. 2014;534(2):383-9.

16. Berthold M, Hand DJ. Intelligent data analysis: an introduction: Springer Science & Business Media; 2003.

17.Ng A. Sparse autoencoder. CS294A Lecture notes. Stanford Univ.[Online]. Available: http://www. stanford. edu/class/cs294a/sparseAutoencoder. pdf; 2011.

18.Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences. 1999;96(12):6745-50.

19.Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer. The lancet. 2002;359(9306):572-7.

20.Komatsu S, Ichikawa D, Takeshita H, Tsujiura M, Morimura R, Nagata H, et al. Circulating microRNAs in plasma of patients with oesophageal squamous cell carcinoma. British journal of cancer. 2011;105(1):104-11.

21.Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer cell. 2002;1(2):203-9.

22. Gashaw I, Grümmer R, Klein-Hitpass L, Dushaj O, Bergmann M, Brehm R, et al. Gene signatures of testicular seminoma with emphasis on expression of ets variant gene 4. Cellular and Molecular Life Sciences CMLS. 2005;62(19-20):2359-68.

23.Aliferis CF, Tsamardinos I, Massion PP, Statnikov AR, Fananapazir N, Hardin DP, editors. Machine Learning Models for Classification of Lung Cancer and Selection of Genomic Markers Using Array Gene Expression Data. FLAIRS Conference; 2003.

24.Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C-H, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences. 2001;98(26):15149-54.