Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 23, Sayı 1, 126-132, 2019

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 23, Issue 1, 126-132, 2019

DOI: 10.19113/sdufenbed.453462

DNA Microarray Gene Expression Data Classification Using SVM, MLP, and RF with Feature Selection Methods Relief and LASSO

Kıvanç GÜÇKIRAN[1], **İsmail CANTÜRK**[2], **Lale ÖZYILMAZ**[3]

[1,2,3]Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Elektronik ve Haberleşme Mühendisliği Bölümü, 34220, İstanbul
[1](ORCID:https://orcid.org/0000-0002-9501-2068)
[2](ORCID:https://orcid.org/0000-0003-0690-1873)
[3](ORCID:https://orcid.org/0000-0001-9720-9852)

(Alınış / Received: 14.08.202018, Kabul / Accepted: 02.04.2019, Online Yayınlanma / Published Online: 24.04.2019)

Keywords
DNA microarray,
Gene expression,
Support vector machines,
Multilayer perceptron,
LASSO,
Relief

Abstract: DNA microarray technology is a novel method to monitor expression levels of a large number of genes simultaneously. These gene expressions can be and are being used to detect various forms of diseases. Using multiple microarray datasets, this paper cross compares two different methods for classification and feature selection. Since individual gene count in microarray data are too many, most informative genes should be selected and used. For this selection, we have tried Relief and LASSO feature selection methods. After picking informative genes from microarray data, classification is performed with Support Vector Machines (SVM), Multilayer Perceptron Networks (MLP) and Random Forest (RF) methods which are widely used in multiple classification tasks. The overall accuracy and training time with LASSO and SVM outperforms most of the approaches proposed.

LASSO ve Relief Özellik Seçimi Yöntemleri ile DVM, ÇKA ve RO Ağ Yapıları Kullanılarak DNA Mikroçip Gen İfadesi Verisetlerinin Sınıflandırılması

Anahtar Kelimeler
DNA mikroçip,
Gen ifadesi,
Destek vektör makinaları,
Çok atmanlı algılayıcı,
LASSO,
Relief

Özet: DNA Mikroçip teknolojisi, çok sayıda gen ifadesinin aynı anda gözlemlenebilmesini sağlayan özgün bir yöntemdir. Günümüzde bu gen ifadeleri bir çok hastalığı teşhis etmek için kullanılmaktadırlar. Bu çalışma iki özellik seçimi ve ağ yapısını çaprazlayarak birden çok verisetinde karşılaştırma yapmaktadır. Mikroçip verisetlerinde her bir örneğin gen sayısı çok sayıda olduğu için, bilgi kazancı en yüksek olan gen seçimi yapılmalıdır. Bu seçim için Relief ve LASSO özellik seçimi yöntemlerini kullandık. En önemli genler örnekten seçildikten sonra Destek Vektör Makinası (DVM), Çok Katmanlı Algılayıcı (ÇKA) ve Rastgele Orman (RO) gibi sıklıkla kullanılan sınıflandırıcılar kullanılarak veri sınıflandırıldı. LASSO özellik seçimi ve DVM daha önceki çalışmaları doğruluk ve eğitim hızı bakımından geride bırakmaktadır.

1. Introduction

DNA microarray data allows us to monitor thousands of different gene expression levels simultaneously [1]. These gene expression levels are used to identify and extract certain information regarding their biological activity. There are various studies which examine the relationship and correlation between gene expression levels and certain types of cancers [2].
Classification of microarray data is not a trivial task. There are multiple approaches being used by the bioinformatics community in order to diagnose and classify the microarray data with the help of machine learning systems. Since each sample of the microarray data consists of thousands of genes to evaluate, feature elimination methods are widely used and adopted [3].

We compare two feature elimination methods, Relief [4] and LASSO [5] with three classifiers, Support Vector Machine, Multilayer Perceptron, and Random Forest. Relief is a highly successful feature elimination method which detects conditional dependencies between features. On the other hand, Least Absolute Shrinkage and Selection Operator (LASSO) is an L1 regularizer with a feature coefficient map. These two feature selection methods are used to prepare the input feature space (DNA microarray data) of the system. First classifier, Support Vector Machine is the de facto classifier for bioinformatics related systems [6]. To compare the results obtained from the Support Vector Machine, we propose two other architectures, Mul-

Corresponding author: kivancguckiran@gmail.com tilayer Perceptron, and Random Forest. There are also certain studies on Multilayer Perceptron usage with DNA microarray data [7], and with Random Forest [8].

We propose simple but powerful approaches to various microarray datasets with binary and non-binary classes using different classification and feature elimination methods. The novelty of this study lies within the combination of the feature selection and classifier methods. Especially SVM classifier with LASSO feature selection achieves the best accuracy and training times compared to other approaches suggested by other studies. We have tested this approach on 16 different datasets. Also, we emphasize not only on the feature selection method but also the accuracy of the classification after feature selection. There are multiple studies on stabilization and the robustness of the feature selection algorithms like [9], [10]. Accuracy evaluation is achieved with Leave-One-Out-Cross-Validation (LOOCV) method, which is the most reliable method to obtain accuracy since there are very few samples in microarray datasets. Section 2 elaborates on materials and methods used in this study. In this section, each dataset is explained and also feature elimination, classification and validation methods and the system proposed are described extensively. Section 3 describes of our simulation and results. This section contains detailed explanation regarding our system. The last section is focused on the conclusion and remarks for future research.

2. Material and Method

This section emphasizes the datasets and methods used by the system thoroughly. In first subsection datasets are explained. Next, the feature selection, classification and cross-validation methods are given in detail and lastly, our novel methodology is introduced.

2.1. Datasets

Total of 22 datasets, which each of them consists of small sample high dimensional DNA microarray input space. In this section, each dataset is named by their author and given in Table 1.

Most of the datasets are for detecting cancer/healthy samples. Also, there are cancer/disease classification datasets. In addition to diseases, there are also 3 microarray datasets which are neither disease nor cancer, like aging and environmental effects on DNA. Datasets can be found as CSV files at https://github.com/kivancguckiran/microarray-data.

2.2. Feature Selection

Feature selection methods are used to reduce the input space of the dataset and remove the unnecessary/unrelated input dimensions. Since microarray data have very high dimensional input space, feature selection is mandatory for these studies. There are two methods handled in our study, Relief, and LASSO.

Relief feature selection algorithm is proposed by Kira and Rendell in 1992 [4]. The algorithm calculates feature value differences with nearest neighbor pairs. Later, these feature scores are used as important values for the system. High feature score means high importance.

LASSO constructs a linear model and penalize the regression coefficients with L1 distance [5]. Most coefficients are reduced to zero and the remaining inputs are selected.

2.3. Classification

Classification is one of the important aspects of machine learning. It enables pattern recognition and likelihood estimation within given datasets. We have examined and implemented two of the widely adopted network architectures, Multilayer Perceptron (MLP) and Support Vector Machine (SVM).

The feedforward multilayer perceptron is a strong classifier which is widely used throughout the machine learning community [11]. MLP contains one or more hidden layers between the input and output layer. All nodes (neurons) are connected with each other between layers. Each connection has a weight which is trained with the backpropagation algorithm.

Random Forest is a classification and regression algorithm proposed By Breiman [12]. The algorithm is based on bootstrap aggregation (bagging), with the extension of feature subset randomization, hence the name, Random Forest. Since its development, the algorithm is used heavily throughout all types of machine learning problems.

Support Vector Machines are the most used classifier among bioinformatics studies. Introduced by Vapnik in 1992 [13], SVM is based on margin maximization and structural risk minimization. Since SVM is very powerful on small sized datasets, it is the state of the art mechanism for microarray data.

2.4. Cross Validation

Cross-validation is an important step to acquire the validity of results. For small sampled datasets like microarray data, it is convenient to use Leave-one-out Cross Validation (LOOCV) for the task [14]. Leave-one-out Cross Validation works by leaving each sample out one at a time and test the system with that sample. Each evaluation is then averaged to compute the overall score.

2.5. Methodology

Our methodology consists of multiple classification tasks. For MLP, we propose 2 hidden layers with 200 and 100 neurons respectively, with Rectified Linear Units (ReLU) as activation functions at each hidden layer [15]. Output layer's activation function is Softmax to overcome classification tasks with multiple classes. As for the backpropagation algorithm, Stochastic Gradient Descent (SGD) with learning rate as 0.005 and momentum as 0.9 is used [16]. Optimizer is selected as Adamax [17]. For batch size as 1, training is achieved with 50 epochs. Weights are initialized via Lecun Uniform distribution [18]. MLP architecture can be seen in Figure **??**.

We have selected estimator count of 100 and a maximum depth of 2 for Random forest algorithm. There was no significant change in classification accuracy regarding these parameters. Estimator count and maximum depth change only affected the training time for the algorithm.

Table 1. Detailed Datasets Table

Dataset	Disease	No. of Samples	No. of Features	No. of Classes
Chin (2006) [19]	Breast Cancer	118	22215	2
Chowdary (2006) [20]	Breast Cancer	104	22283	2
Gravier (2010) [21]	Breast Cancer	168	2905	2
Sorlie (2001) [22]	Breast Cancer	85	456	5
West (2001) [23]	Breast Cancer	49	7129	2
Pomeroy (2002) [24]	CNSET	60	7128	2
Burcyznski (2006) [25]	Crohn's Disease	127	22283	3
Alon (1999) [26]	Colon Caner	62	2000	2
Sun (2006) [27]	Glioma	180	54613	4
Borovecki (2005) [28]	Huntington's Disease	31	22283	2
Chiaretti (2004) [29]	Leukemia	128	12625	2
Golub (1999) [30]	Leukemia	72	7129	2
Yeoh (2002) [31]	Leukemia	248	12625	6
Gordon (2002) [32]	Lung Cancer	181	12533	2
Shipp (2002) [33]	Lymphoma	77	6817	2
Tian (2003) [34]	Myeloma	173	12625	2
Singh (2002) [35]	Prostate Cancer	102	12600	2
Nakayama (2007) [36]	Sarcoma	105	22283	10
Khan (2001) [37]	SRBCT	63	2308	4
Christensen (2009) [38]	N/A	217	1413	3
Su (2002) [39]	N/A	102	5565	4
Subramanian (2005) [40]	N/A	50	10100	2

We have used linear kernel for SVM with L2 penalty, the loss function is calculated via squared hinge loss and we have selected tolerance of the stopping criteria as $1e^{-4}$ and the maximum iteration count as 1000.

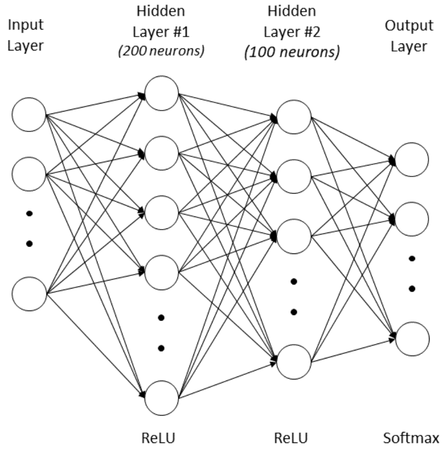

Figure 1. MLP Architecture

Our steps are as follows, first, we apply the LASSO feature selection algorithm to the datasets and get the selected features. We have selected as the alpha parameter for LASSO as 0.001. Feature count at this step is noted for the related dataset. Then, we use MLP and SVM classifiers to classify and validate the results using LOOCV to calculate their scores. After this step, we apply the Relief feature selection algorithm to the datasets and select the most important features. Selected feature count is taken from LASSO to equally compare their performances. Like the step before, SVM and MLP networks are used to classify and LOOCV to validate. The number of features selected for each dataset can be seen in Table 2. Figure **??** represents our methodology in a diagram. The feature selection and classification is achieved with the help of Python and frameworks like NumPy, SciPy, Pandas, Keras and Matplotlib. The code for training and testing is open-sourced and can be found at https://github.com/kivancguckiran/microarray-classification.

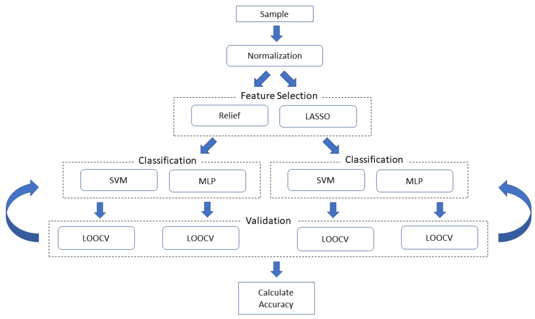

Figure 2. Methodology

3. Results

Using LOOCV, each sample is left out and other samples are used for training. After the training, the sample we left out is validated against the network and the result is noted. After the training and tests, the accuracy results are very promising. Results can be seen from Table 2.

The training times must be noted since MLP training takes way much longer time. SVM is nearly 500 times faster. We can also derive from the results that, LASSO feature

Table 2. Results

Dataset	Features	MLP		SVM		RF	
		LASSO	Relief	LASSO	Relief	Lasso	Relief
Chin	112	**100%**	85.6%	**100%**	85.6%	90.7%	89.0%
Chowdary	83	**100%**	99.0%	**100%**	99.0%	98.1%	96.2%
Gravier	128	96.4%	79.8%	**98.2%**	79.8%	74.4%	71.4%
Sorlie	74	88.2%	**92.9%**	84.8%	88.2%	62.4%	76.5%
West	51	**98.0%**	87.6%	94.2%	89.8%	95.9%	85.7%
Pomeroy	56	**100%**	85%	**100%**	85%	66.7%	75.0%
Burcyznski	122	**97.6%**	91.3%	93.5%	89.8%	75.6%	80.3%
Alon	58	96.8%	85.5%	**98.4%**	83.9%	83.9%	85.5%
Sun	181	**89.4%**	70%	83.3%	67.8%	60.0%	69.4%
Borovecki	30	**100%**	96.8%	**100%**	**100%**	**100%**	96.8%
Chiaretti	123	95.3%	86.7%	**96.1%**	88.3%	81.3%	79.7%
Golub	64	**100%**	95.8%	**100%**	97.2%	98.6%	97.2%
Yeoh	231	**98.8%**	98.0%	98.0%	98.4%	84.3%	74.6%
Gordon	97	**100%**	98.9%	**100%**	98.9%	99.4%	98.9%
Shipp	66	**100%**	96.1%	**100%**	98.7%	93.5%	92.2%
Tian	158	**100%**	79.2%	**100%**	82.7%	79.2%	78.6%
Singh	84	99.0%	95.1%	**100%**	95.1%	94.1%	93.1%
Nakayama	129	68.6%	**74.3%**	69.7%	64.8%	58.1%	53.3%
Khan	56	**100%**	**100%**	**100%**	**100%**	**100%**	85.7%
Christensen	72	**100%**	**100%**	**100%**	**100%**	**100%**	98.6%
Su	90	**100%**	99.0%	**100%**	**100%**	99.0%	96.1%
Subramanian	46	**100%**	94%	**100%**	96%	88.0%	90.0%
Average		**96.7%**	90.5%	**96.2%**	90.4%	**85.6%**	84.7%

selection significantly outperforms Relief feature selection. From the network perspective, SVM is overwhelmingly faster at training, but for accuracy, MLP is slightly better. Random Forest's training time is faster than MLP and slower than SVM, but in terms of accuracy, unfortunately, it was the least successful among classifiers used in this study.

There are a great number of studies on DNA microarray classification. We have tried to pick approaches which contain state-of-the-art evaluations on multiple datasets to achieve a fair comparison in terms of accuracy. In addition to this, it must be noted that the cross-validation methods might not be LOOCV in these studies. For example, in [41], Arias-Michel uses 5 fold cross-validation and in [42], Huertas derives cross-validation strategy from Logistic Regression.

We compare our proposed approach, LASSO and SVM, to other studies, and notice that our approach is significantly better. There are various works within this area of research and there are multiple approaches to compare. Since most of the studies focused mostly on one or two datasets, we have tried the pick the ones that generalize well and contains multiple datasets to compare. Accuracies reported from other studies for comparison were the best results within their studies. The comparison can be seen in Table 3. "Ours" column represents the Lasso and SVM approach. We have paid too much attention on not to cherry pick the results for comparison. For example, we have taken the best results from Arias-Michel's study [41], even though the best results were not from the same algorithm. But on our side, we have picked LASSO and SVM strictly for the comparison.

4. Discussion and Conclusion

In this study, we propose a network with LASSO and SVM to classify microarray datasets with few samples and very high dimensional input space. The method is proved to be a reliable classification system compared to other approaches. Although in terms of accuracy, MLP is slightly better than SVM, the training speed is overwhelmingly slower. In addition to this, MLP accuracy heavily depends on the initial values. If we take the training speed and the uncertainty of the training into consideration, it can be argued that LASSO and SVM is the viable solution to this problem. Using SVM as a classifier on a DNA microarray classification problem is widely accepted among literature, but none of the previous works combined LASSO feature selection with SVM.

Unlike other linear regularization models, LASSO can further decrease coefficients to zero using L1 regularization. Since microarray data has a great number of features, canceling out coefficients reduces variance greatly. This way, LASSO also prevents model to overfit to the data. It is observed that these advantages make LASSO a viable approach for DNA microarray data compared to other feature selection methods. As for the classification method, SVM works incredibly well with small datasets using hyperplanes to separate different classes. This makes SVM an optimal approach for microarray data since most of the datasets have around 100 samples.

Table 3. Comparison

Dataset	Ari. [41]	Hue.[42]	Phu. [43]	Mun. [44]	Le. [45]	Ours
Alon	-	82.1%	88.7%	91.1%	82.5%	**98.4%**
Gravier	76.2%	75.0%	79.8%	-	-	**98.2%**
Chow.	95.2%	97.3%	98.1%	-	-	**100%**
Tian	-	78.7%	-	-	-	**100%**
Golub	-	-	**100%**	98.4%	-	**100%**
Pomer.	63.3%	-	-	92.1%	-	**100%**
Shipp	92.3%	-	58.6%	-	-	**100%**
Gordon	99.4%	-	**100%**	99.9%	99.3%	**100%**
Singh	99.5%	-	-	97.2%	-	**100%**
Chin	88.0%	-	-	-	-	**100%**
Burc.	86.5%	-	-	-	-	**93.5%**
Chia.	-	-	85.2%	-	-	**96.1%**
Khan	-	-	-	-	99.6%	**100%**
Yeoh	-	-	-	-	97.1%	**98.0%**
Nak.	-	-	-	-	**89.9%**	69.7%
Sun	-	-	-	-	72.3%	**83.3%**

We propose this system as a general approach to classify DNA microarray datasets. We have tried to generalize the approach by increasing the dataset variety. Since all of the classifiers' overall performance is pretty well, it can be argued that the challenge microarray datasets pose is the feature selection. It should be noted that Nakayama [36] dataset with the soft tissue sarcoma disease has the highest number of classes and the least accuracy. It can further be studied with more data if any microarray dataset with sarcoma disease comes to light or with data augmentation methods in the future.

References

[1] Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270(5235), 467-470.

[2] Alizadeh, Ash & B Eisen, Michael & Davis, Richard & Ma, Chi & S Lossos, Izidore & Rosenwald, Andreas & C Boldrick, Jennifer & Sabet, Hajeer & Tran, Truc & Yu, Xin. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 403. 503-511.

[3] Hira, Z. M., & Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Advances in Bioinformatics, 2015, 198363.

[4] Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Machine Learning Proceedings 1992 (pp. 249-256).

[5] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[6] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Ares, M., & Haussler, D. (1999). Support vector machine classification of microarray gene expression data. University of California, Santa Cruz, Technical Report UCSC-CRL-99-09.

[7] Rafii, F., Kbir, M. H. A., & Hassani, B. D. R. (2015, November). MLP network for lung cancer presence prediction based on microarray data. In Complex Systems (WCCS), 2015 Third World Conference on (pp. 1-6). IEEE.

[8] Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1), 3.

[9] Drotár, P., Gazda, J., & Smékal, Z. (2015). An experimental comparison of feature selection methods on two-class biomedical datasets. Computers in biology and medicine, 66, 1-10.

[10] Gutkin, M., Shamir, R., & Dror, G. (2009). SlimPLS: a method for feature selection in gene expression-based disease classification. PloS one, 4(7), e6416.

[11] Lippmann, R. (1987). An introduction to computing with neural nets. IEEE Assp magazine, 4(2), 4-22.

[12] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[13] Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media.

[14] Radmacher, M. D., McShane, L. M., & Simon, R. (2002). A paradigm for class prediction using gene expression profiles. Journal of Computational Biology, 9(3), 505-511.

[15] Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 807-814).

[16] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010 (pp. 177-186). Physica-Verlag HD.

[17] Kingma, D. P., & Ba, J. (2014). Adam: A

method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[18] LeCun, Y., Huang, F. J., & Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on (Vol. 2, pp. II-104). IEEE.

[19] Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., ... & Chen, F. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer cell, 10(6), 529-541.

[20] Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., ... & Mazumder, A. (2006). Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. The journal of molecular diagnostics, 8(1), 31-39.

[21] Gravier, E., Pierron, G., VincentSalomon, A., Gruel, N., Raynal, V., Savignoni, A., ... & Fourquet, A. (2010). A prognostic DNA signature for T1T2 nodenegative breast cancer patients. Genes, chromosomes and cancer, 49(12), 1125-1134.

[22] Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., ... & Thorsen, T. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences, 98(19), 10869-10874.

[23] West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., ... & Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. Proceedings of the National Academy of Sciences, 98(20), 11462-11467.

[24] Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., ... & Allen, J. C. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, 415(6870), 436.

[25] Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., ... & Spinelli, W. (2006). Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. The journal of molecular diagnostics, 8(1), 51-61.

[26] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). 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, 96(12), 6745-6750.

[27] Sun, L., Hui, A. M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., ... & Rosenblum, M. (2006). Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer cell, 9(4), 287-300.

[28] Borovecki, F., Lovrecic, L., Zhou, J., Jeong, H., Then, F., Rosas, H. D., ... & Krainc, D. (2005). Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. Proceedings of the National Academy of Sciences, 102(31), 11023-11028.

[29] Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., ... & Foa, R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood, 103(7), 2771-2778.

[30] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

[31] Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., ... & Cheng, C. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer cell, 1(2), 133-143.

[32] Gordon, G. J., Jensen, R. V., Hsiao, L. L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., ... & Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer research, 62(17), 4963-4967.

[33] Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., ... & Ray, T. S. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine, 8(1), 68.

[34] Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., & Shaughnessy Jr, J. D. (2003). The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. New England Journal of Medicine, 349(26), 2483-2494.

[35] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., ... & Lander, E. S. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer cell, 1(2), 203-209.

[36] Nakayama, R., Nemoto, T., Takahashi, H., Ohta, T., Kawai, A., Seki, K., ... & Hasegawa, T. (2007). Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma. Modern pathology, 20(7), 749.

[37] Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., ... & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine, 7(6), 673.

[38] Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., ... & Sugarbaker, D. J. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS genetics, 5(8), e1000602.

[39] Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y.,

Walker, J. R., Wiltshire, T., ... & Patapoutian, A. (2002). Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences, 99(7), 4465-4470.

[40] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545-15550.

[41] Arias-Michel, R., García-Torres, M., Schaerer, C. E., & Divina, F. (2015, September). Feature selection via approximated Markov blankets using the CFS method. In Data Mining with Industrial Applications (DMIA), 2015 International Workshop on (pp. 38-43). IEEE.

[42] Huertas, C., & Juarez-Ramirez, R. (2016). Automatic Threshold Search for Heat Map Based Feature Selection: A Cancer Dataset Analysis. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 10(7), 1341-1347.

[43] Huynh, P. H., Nguyen, V. H., & Do, T. N. (2018, December). Random ensemble oblique decision stumps for classifying gene expression data. In Proceedings of the Ninth International Symposium on Information and Communication Technology (pp. 137-144). ACM.

[44] Mundra, P. A., & Rajapakse, J. C. (2010). Gene and sample selection for cancer classification with support vectors based t-statistic. Neurocomputing, 73(13-15), 2353-2362.

[45] Le Thi, H. A., & Phan, D. N. (2017). DC programming and DCA for sparse Fisher linear discriminant analysis. Neural Computing and Applications, 28(9), 2809-2822.