**Original Article** *Iran J Ped Hematol Oncol. 2023, Vol 13, No 1, 13-21*

# Prediction of blood cancer using leukemia gene expression data and sparsity-based gene selection methods

Sanaz Mehrabani[1], Morteza Zangeneh Soroush[2], Negin Kheiri[3], Razieh Sheikhpour[4], Mahshid Bahrami[5*]

1. Non-Communicable Pediatric Diseases Research Center, Health Research Institute, Babol University of Medical Sciences, Babol, Iran
2. Department of Biomedical Engineering, Science and Research branch, Islamic Azad University, Tehran, Iran
3. Shiraz University of Medical Sciences, Shiraz, Iran
4. Department of Computer Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran
5. Department of Radiology, Isfahan University of Medical Sciences, Isfahan, Iran

*Corresponding author: Dr Mahshid Bahrami, Department of Radiology, Isfahan University of Medical Sciences, Isfahan, Iran. Email: mahshidbahrami273@yahoo.com. ORCID ID: 0000-0002-0799-8059

**Received:** 19 June 2022 **Accepted:** 13 August 2022

**Abstract**

**Background:** DNA microarray is a useful technology that simultaneously assesses the expression of thousands of genes. It can be utilized for the detection of cancer types and cancer biomarkers. This study aimed to predict blood cancer using leukemia gene expression data and a robust $\ell_{2,p}$-norm sparsity-based gene selection method.

**Materials and Methods**: In this descriptive study, the microarray gene expression data of 72 patients with acute myeloid leukemia (AML) and lymphoblastic leukemia (ALL) was used. To remove the redundant genes and identify the most important genes in the prediction of AML and ALL, a robust $\ell_{2,p}$-norm ($0 < p \leq 1$) sparsity-based gene selection method was applied, in which the parameter p method was implemented from 1/4, 1/2, 3/4 and 1. Then, the most important genes were used by the random forest (RF) and support vector machine (SVM) classifiers for prediction of AML and ALL.

**Results:** The RF and SVM classifiers correctly classified all AML and ALL samples. The RF classifier obtained the performance of 100% using 10 genes selected by the $\ell_{2,1/2}$-norm and $\ell_{2,1}$-norm sparsity-based gene selection methods. Moreover, the SVM classifier obtained a performance of 100% using 10 genes selected by the $\ell_{2,1/2}$-norm method. Seven common genes were identified by all four values of parameter p in the $\ell_{2,p}$-norm method as the most important genes in the classification of AML and ALL, and the gene with the description "PRTN3 Proteinase 3 (serine proteinase, neutrophil, Wegener granulomatosis autoantigen" was identified as the most important gene.

**Conclusion:** The results obtained in this study indicated that the prediction of blood cancer from leukemia microarray gene expression data can be carried out using the robust $\ell_{2,p}$-norm sparsity-based gene selection method and classification algorithms. It can be useful to examine the expression level of the genes identified by this study to predict leukemia.

**Keywords:** Gene expression data, Gene Selection, Acute myeloid leukemia, Acute lymphoblastic leukemia

## Introduction

Leukemia is an amalgam of cancers that occurs due to the malignancy of the blood and bone marrow elements (1). On the other hand, it is abnormal white blood cells with incomplete development called blasts or leukemia cells (2). It is the most common cancer in children which is caused by infection, ionizing radiation, and genetic factors (3). Acute leukemia is a heterogeneous group of diseases categorized into lymphoid and myeloid leukemias based on the immunologic markers (2). The most common frequent types of leukemia among children are acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) (4). AML (5,6) and ALL are the cancer of the myeloid, and lymphoid line of blood cells, respectively (4,5). Abnormal cells are rapidly reproduced in the bone marrow and blood, which leads to the dysfunction of normal blood cells in AML and ALL (5). Diagnosis of ALL from AML concerning treatment and prognosis is important. One



[ DOI: 10.18502/ijpho.v13i1.11629 ]

of the precise methods for diagnosis of ALL and AML is the use of genetic information and the patient's DNA (3). DNA microarray technology is a possible method for the diagnosis of ALL and AML. Moreover, it can assess the expression of thousands of genes, simultaneously (7). The data of microarray gene expression are used to detect cancer biomarkers, cancer diagnosis, and biomarkers or gene signatures (8–10). The number of genes in microarray data is larger than the sample number, leading to challenges in the analysis of microarray data (8–10). Most of the genes in the microarray data are redundant, and a few relevant genes may be useful for cancer diagnosis and appropriate treatment selection in clinical management. Therefore, one of the most important steps in the analysis of microarray data is to reduce the number of genes and choose the proper genes to classify cancer, leading to the reduced processing time of classification and misclassification rate (9,11). To analyze the microarray data, gene selection methods and machine learning algorithms can be used to reduce the number of genes, remove the redundant genes, and identify the relevant ones (3). Gene selection methods can be divided into the filter, wrapper, and embedded methods. Filter methods are rapid and simple which is appropriate for microarray data with thousands of genes. In the filter methods, the genes are ranked according to certain characteristics independent of the classifiers. Wrappers apply some criteria for selecting several genes with the best performance for a specified classifier. Although these methods have good performance, their computational cost is high. In the embedded gene selection methods, gene selection and classification are done simultaneously (12). In the classical filter gene selection methods, the correlation among genes is ignored and the importance of each gene is individually evaluated. The sparsity-based gene selection methods use joint sparse models to consider the correlated information among different features in the gene selection process (3,13). In this study, a graph-based $\ell_{2,p}$-norm sparsity-based gene selection method was applied to identify the most important genes of leukemia gene expression data in the prediction of AML and ALL. The $\ell_{2,p}$-norm sparsity-based gene selection method is robust to outliers and considers the distribution of leukemia data in the gene selection process. Then random forest (RF) and support vector machine (SVM) classifiers were used to predict AML and ALL using the selected genes.

## Materials and Methods

### Data

This study is descriptive, and the leukemia data was obtained from the microarray gene expression data collected from the bone marrow of leukemia cancer patients presented by Golub et al. (14). In the leukemia gene expression data, there is information about 72 patients are divided into 23 AML and 49 ALL samples, which include the expression of 7129 genes. To classify the leukemia data into the AML and ALL, the gene expression data should be split into the training set for model construction and the test set for model evaluation. In this study, the training set contains the gene expression data of 38 leukemia patients (25 ALL and 13 AML) and the test set contains the gene expression data of 34 leukemia patients (24 ALL and 10 AML), respectively.

### *The graph-based $\ell_{2,p}$-norm sparsity-based gene selection method*

To consider the importance of genes in the prediction of AML and ALL, a robust graph-based $\ell_{2,p}$-norm sparsity-based gene selection method called $G^S3FS$ was used which considers the distribution and geometry structure of leukemia data in the gene selection process. The gene selection method applies a loss function and regularization based on $\ell_{2,p}$-norm ($0 < p \leq 1$), which is robust to outlier and makes



[ DOI: 10.18502/ijpho.v13i1.11629 ]

the problem convex when $p$=1 and non-convex when $0 < p < 1$. The formulation of the $\ell_{2,p}$-norm gene selection method is defined as (13):

$$\arg\min_{W,b} Tr(W^T X L X^T W) + \|X^T W + 1_n b^T - Y\|_{2,p}^p + \lambda \|W\|_{2,p}^p, p \in (0,1]$$

(1)

In Eq. (1), *Tr* (.) is the trace operator, *X* is the leukemia data, *Y* is the leukemia types such as the AML and ALL, *W* is a projection matrix utilized for gene selection, *b* is the bias term, 1 is a column vector in which all elements are 1, and *n* is the number of training leukemia data. $\lambda$ indicates the regularization parameters, which is set to 1 in this study, and *L* is the graph Laplacian. To calculate the graph Laplacian, a graph *S* with *n* nodes is created in which each node indicates each leukemia sample. In this graph, close leukemia samples are connected, and the weight matrix of this graph between an *i*th sample and the *j*th sample is calculated as:

$$S_{i,j} = \begin{cases} 1 & \text{if } x_i \in KNN(x_j) \text{ or } x_j \in KNN(x_i) \\ 0 & otherwise \end{cases}$$

(2)

where $x_i$ is the *i*th sample in leukemia data. If $x_i$ *is in among the K nearest neighbors of* $x_i$ or $x_j$ *is in among the K nearest neighbors of* $x_j$, the $S_{i,j}$ will be one. After the construction of the graph, the graph Laplacian is calculated as $L = D - S$, where $D$ is a diagonal matrix computed as $D_{ii} = \sum_j S_{ij}$.

The objective function in Eq. (1) can use a predicted label matrix as $F = [f_1, f_2, \ldots, f_n]^T \in R^{n \times c}$ using the transductive classification algorithm in (15), where *c* is the number of the leukemia types such as AML and ALL. According to (16), *F* was defined as Eq. (3) to be smooth on the graph model *S* and be consistent with the labels of the training data.

$$\arg\min_{F} Tr(F^T L F) + Tr(F - Y)^T U (F - Y)$$

(3)

In Eq. (3), $U \in R^{n \times n}$ is a diagonal decision rule matrix, in which $U_{ii} = \infty$ if $x_i$ is labeled training data and $U_{ii}$=1 otherwise.

By combining Eq. (3) into Eq. (1), the graph-based $\ell_{2,p}$-norm sparsity-based gene selection method G$^S$3FS is defined as:

$$\arg\min_{F,W,b} Tr(F^T L F) + Tr\big((F - Y)^T U (F - Y)\big) + \|X^T W + 1_n b^T - F\|_{2,p}^p + \lambda \|W\|_{2,p}^p, \; p \in (0,1]$$

(4)

The $\ell_{2,p}$-norm gene selection method in Eq. (4) uses the graph Laplacian of leukemia data to consider the distribution and geometry structure of data, utilizes the $\ell_{2,p}$-norm on loss function to make the objective robust to the outlier, and the $\ell_{2,p}$-norm on regularization to make the objective appropriate for a selection of genes and consider the correlation between genes in the gene selection process.

**Random forest**

Random forest (RF) is a classification method that creates a large number of decision trees in the training process and classifies the leukemia data based on the class selected by most trees. Each tree in the random forest predicts AML or ALL class, and the class with the most tree becomes the prediction of RF (17).

**SVM**

Support vector machine (SVM) is a powerful classification algorithm that searches a boundary called a hyperplane which separates and classifies the leukemia data into AML and ALL. The main purpose of the SVM is to search the hyperplane with the maximum margin that provides the maximum separation between AML and ALL. In many real-world applications, the relationships are nonlinear and the classes are not linearly separable, and SVM can map the feature space to higher dimensions using a kernel trick, which may result in linear



[ DOI: 10.18502/ijpho.v13i1.11629 ]

relationships. Thus similar classes can be separated linearly (18).

**Ethical Considerations**

This study was approved by the Ethical Committee of Shahid Sadoughi University of Medical Sciences (IR.SSU.MEDICINE.REC.1401.143.).

## Results

To predict the AML and ALL based on the leukemia gene expression data, we initially performed several experiments to identify the importance of each gene using the $\ell_{2,p}$-norm sparsity-based gene selection method (GS3FS) in Eq. (1). In the experiments, parameter p in the $\ell_{2,p}$-norm sparsity-based gene selection method was implemented from {0.25, 0.5, 0.75, 1}, the value of parameter λ in Eq. (1) was set to 1, and the value of parameter k which indicates the number of nearest neighbors for computing the graph Laplacian was set to 5. The performance of the $\ell_{2,p}$-norm sparsity-based gene selection method was also compared with SFSS (19,20), a sparse gene selection method that applies the $\ell_2$-norm based loss function and $\ell_{2,1}$-norm regularization, and FSLG (21) which is a sparse gene selection method based on $\ell_2$-norm loss function and $\ell_{2,1/2}$-norm regularization. Then, the most important and relevant genes are selected and given to the SVM and RF classifiers to predict the leukemia type and classify the data into AML and ALL classes. The experiments performed by RF are repeated 10 times and the best results are reported. The number of trees in the RF classifier is set to 20. The linear kernel is used in the SVM classifier. The criteria used to assess the efficiency of the SVM classifier constructed by the genes selected by the $\ell_{2,p}$-norm sparsity-based gene selection method were accuracy, sensitivity, and specificity. The accuracy criterion indicates the percentage of leukemia samples that are correctly classified into AML and ALL. The sensitivity criterion shows the percentage of leukemia data that are correctly classified into AML class, and the specificity demonstrates the percentage of leukemia data that are correctly classified into ALL class. The performance of RF and SVM classifiers for the classification of AML and ALL using the 10 and 15 genes selected by the $\ell_{2,p}$-norm sparsity-based gene selection method are shown in Tables I and II, respectively. As shown in Tables I and II, the RF classifier obtained the performance of 100% and correctly classified all AML and ALL samples using 10 genes selected by the GS$^3$FS method based on $\ell_{2,1/2}$-norm $\ell_{2,1}$-norm, and 15 genes selected by the GS$^3$FS method based on $\ell_{2,1/4}$-norm, $\ell_{2,1/2}$-norm, and $\ell_{2,1}$-norm. Moreover, the performance of 100% was obtained by the SVM classifier using 10 genes selected by the GS$^3$FS method based on $\ell_{2,1/2}$-norm. The reason for the high efficiency of the $\ell_{2,p}$-norm ($0 < p \leq 1$) gene selection method in identifying the most important genes was the use of distribution and geometry structure of all leukemia samples and the $\ell_{2,p}$-norm minimization on both regularization and loss function. The results of Tables I and II also demonstrated that RF and SVM classifiers using the genes selected by the $\ell_{2,p}$-norm sparsity-based gene selection method (GS$^3$FS) had better performance than the genes selected by SFSS and FSLG methods. This is because the GS$^3$FS gene selection method applies the $\ell_{2,p}$-norm on regularization and loss function which makes the method robust to the outlier and selects the most relevant genes in the prediction of AML and ALL. To investigate the important and relevant genes of leukemia gene expression data, we identified common genes with high rank selected by the $\ell_{2,p}$-norm gene selection method using all four values of parameter *p*, which was demonstrated in Table III.



[ DOI: 10.18502/ijpho.v13i1.11629 ]

*Table I: The performance of RF and SVM classifiers for the classification of AML and ALL using the 10 genes selected by different sparsity-based gene selection methods*

| Classifier | Gen selection method | *Accuracy* | *Sensitivity* | *Specificity* |
|---|---|---|---|---|
| **RF** | $G^S3FS$ ($\ell_{2,1/4}$) | 97.06 | 90 | 100 |
| | $G^S3FS$ ($\ell_{2,1/2}$) | 100 | 100 | 100 |
| | $G^S3FS$ ($\ell_{2,3/4}$) | 91.18 | 70 | 100 |
| | $G^S3FS$ ($\ell_{2,1}$) | 100 | 100 | 100 |
| | SFSS | 94.12 | 100 | 91.67 |
| | FSLG | 79.41 | 30 | 100 |
| **SVM** | $G^S3FS$ ($\ell_{2,1/4}$) | 97.06 | 100 | 95.83 |
| | $G^S3FS$ ($\ell_{2,1/2}$) | 100 | 100 | 100 |
| | $G^S3FS$ ($\ell_{2,3/4}$) | 91.18 | 100 | 87.50 |
| | $G^S3FS$ ($\ell_{2,1}$) | 44.12 | 100 | 20.83 |
| | SFSS | 94.12 | 100 | 91.67 |
| | FSLG | 20.59 | 100 | 25 |

*Table II: The performance of RF and SVM classifiers for the classification of AML and ALL using the 15 genes selected by different sparsity-based gene selection methods*

| Classifier | Gen selection method | *Accuracy* | *Sensitivity* | *Specificity* |
|---|---|---|---|---|
| **RF** | $G^S3FS$ ($\ell_{2,1/4}$) | 100 | 100 | 100 |
| | $G^S3FS$ ($\ell_{2,1/2}$) | 100 | 100 | 100 |
| | $G^S3FS$ ($\ell_{2,3/4}$) | 97.06 | 90 | 100 |
| | $G^S3FS$ ($\ell_{2,1}$) | 100 | 100 | 100 |
| | SFSS | 91.18 | 100 | 87.50 |
| | FSLG | 82.35 | 60 | 91.67 |
| **SVM** | $G^S3FS$ ($\ell_{2,1/4}$) | 47.06 | 100 | 25 |
| | $G^S3FS$ ($\ell_{2,1/2}$) | 97.06 | 100 | 95.83 |
| | $G^S3FS$ ($\ell_{2,3/4}$) | 97.06 | 100 | 95.83 |
| | $G^S3FS$ ($\ell_{2,1}$) | 94.12 | 100 | 91.67 |
| | SFSS | 38.24 | 100 | 12.5 |
| | FSLG | 76.47 | 20 | 100 |

*Table III. The common genes with high rank selected by the $\ell_{2,p}$-norm gene selection method using all four values of parameter p on leukemia gene expression data*

| Gene Number | Accession | Description |
|---|---|---|
| **4279** | X55668_at | PRTN3 Proteinase 3 (serine proteinase, neutrophil, Wegener granulomatosis autoantigen |
| **6277** | M30703_s_at | Amphiregulin (AR) gene |
| **5954** | Y00339_s_at | CA2 Carbonic anhydrase II |
| **6308** | M57731_s_at | GRO2 oncogene |
| **5972** | X57579_s_at | GB DEF = Activin beta-A subunit (exon 2) |
| **5599** | D28235_s_at | Cyclooxygenase-2 (hCox-2) gene |
| **1745** | M16038_at | LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog |



[ DOI: 10.18502/ijpho.v13i1.11629 ]

## Discussion

In this study, we utilized the gene expression data of leukemia to identify the importance of each gene by the $\ell_{2,p}$-norm sparsity-based gene selection method and select the most important genes for classification into AML and ALL by the RF and SVM classifiers. The RF classifier correctly classified all AML and ALL samples and achieved the performance of 100% using 10 genes selected by the GS$^3$FS method based on $\ell_{2,1/2}$-norm and $\ell_{2,1}$-norm. The best performance of RF in the classification of AML and ALL using 15 genes was 100%, and it was obtained by the G$^S$3FS method based on $\ell_{2,1/4}$-norm, $\ell_{2,1/2}$-norm, and $\ell_{2,1}$-norm. The best classification performance of the SVM classifier using 10 genes was 100%, and it was obtained by the G$^S$3FS method based on $\ell_{2,1/2}$-norm. SVM classifier using 15 genes obtained an accuracy of 97.06% by the G$^S$3FS method based on $\ell_{2,1/2}$-norm, $\ell_{2,3/4}$-norm. Arunkumar and Ramakrishnan (22) proposed a customized similarity measure using a fuzzy rough quick reduct algorithm for gene selection. In the first step, they used the Information Gain based entropy for dimensionality reduction, and in the second step, they employed the proposed fuzzy rough quick reduce method that defines a customized similarity measure to select the minimum number of relevant genes and remove the redundant genes. The fuzzy method achieved a classification accuracy of 90.28% on a raw dataset of leukemia gene expression data, and 97.22% on dimensionality-reduced leukemia gene expression data. Kumar and Halder (23) proposed an ensemble-based active learning using a fuzzy-rough algorithm to classify cancer using microarray gene expression data. The proposed fuzzy-rough algorithm can achieve good classification accuracy with the limited training samples and deal with the uncertainty, overlap, and indiscernibility in the gene expression data. They obtained an accuracy of 97.72 on the leukemia data using their proposed ensemble-based learning algorithm. They also obtained an accuracy of 93.62% utilizing the ALFKNN (24) method which applies the active learning strategy for the selection of the most confusing samples from the unlabeled microarray gene expression data. ALFNKK method selects the most confusing samples and classifies the test samples using the fuzzy $k$-nearest neighbor method. Moreover, they used the ALRFC (25) method, a rough fuzzy algorithm for the selection of the most confusing samples from the pool of unlabeled samples which classifies the test samples with the fuzzy membership value of the test samples for belonging to a particular class, and obtained the accuracy of 98.88% on leukemia data in the diagnosis of ALL and AML. Sun et al. (26) proposed a neighborhood rough set gene selection method (GSFSJNE), which uses the entropy measure-based gene selection with Fisher score for tumor classification. The method can deal with real-value data whilst maintaining the original information of gene classification. They applied the Fisher score for dimension reduction and irrelevant genes elimination. Then, they considered some neighborhood entropy-based uncertainty criteria to handle the uncertainty and noise of gene expression data, derive some of their properties, and establish the relationships among the criteria. Finally, they presented a joint neighborhood entropy-based gene selection algorithm with the Fisher score to select the relevant genes from gene expression data. They used the GSFSJNE gene selection method, selected four important genes from leukemia gene expression data, and classified the genes selected by the GSFSJNE method using SVM, KNN, and C4.5 classifiers with the accuracy of 88.89%, 91.98%, and 77.64%, respectively. Sheikhpour et al. (3) identified the important genes of leukemia data by a sparse gene selection method which uses a quadratic loss function and a regularization based on $\ell_{2,1}$-norm. Then,



[ DOI: 10.18502/ijpho.v13i1.11629 ]

they used SVM, KNN, Gaussian kernel density estimation-based classifier (GKDEC), and linear discriminant classifier (LDC) classifiers using the selected genes to classify AML and ALL. SVM, GKDEC, KNN, and LDC achieved an accuracy of 94.12%, 100%, 91.18%, and 100% on the classification of AML and ALL samples, respectively. Abdullah et al. (27) utilized an ensemble classifier called EOD classifier to predict the leukemia subtypes from microarray gene expression data. The classification accuracy of the EOD classifier in the prediction of AML and ALL was 96.66% using 10-fold cross-validation. Abdullah et al. (27) obtained an accuracy of 98.33%, 90.83%, 95%,94.16%, and 96.66% using naïve Bayes (NB), k-nearest neighbor (KNN), SVM, multi-layer perceptron (MLP) and RF classifiers, respectively. Kumar Dey and Islam (28) predicted AML and ALL from leukemia gene expression data using three machine-learning algorithms. Initially, they applied principal component analysis (PCA) to reduce the dimensionality of leukemia data, and then utilized artificial neural networks (ANN), RF, and XGBoost classifiers to predict the type of leukemia gene expression data. The classification accuracy of ANN, RF, and XGBoost was 92.3%, 80.8%, and 92.3%, respectively. Maria and Thirupathi (29) carried out an analysis of different classification algorithms on leukemia gene expression Data. They used KNN, SVM, NB, logistic regression, RF, and ANN algorithms for the classification of leukemia samples into AML and ALL. They obtained an accuracy of 88.2% and 94.1%. 91.2%, 100%, 91.2%, and 85.3 using KNN, SVM, NB logistic regression, RF, and ANN, respectively. Li et al. (30) applied the weighted gene co-expression networks to split the genes of leukemia data into groups and presented a regularized multinomial regression with overlapping group lasso penalty (MROGL) to select gene groups and carry out multi-classification. MROGL method obtained an accuracy of 95.8% using 612 genes of leukemia data. Wu et al. (31) proposed a method that combines manifold learning and Gaussian process classification to classify gene expression data. They used Isometric feature mapping for dimension reduction of data and manifold feature extraction, and the Gaussian process for gene expression classification. They achieved an accuracy of 98.61% using their proposed method and an accuracy of 97.22% using the SVM classifier and SLIE gene selection method on leukemia gene expression data.

## Conclusion

The results obtained in this study demonstrated that the types of leukemia can be predicted with high accuracy using the genes identified by the robust $\ell_{2,p}$-norm ($0 < p \leq 1$) sparsity-based gene selection method and the machine learning algorithms. The robust $\ell_{2,p}$-norm sparsity-based gene selection method with different values of parameter p considered the distribution of leukemia data to identify the most important genes. The examination and study of the expression level of the selected genes can be useful in the prediction of leukemia.

## Conflict of interest

The authors declare no conflict of interest.

## References

1. Chapla U, Chapalamadugu U, Ojochenemi DA, Chatakonda R. Leukemia-brief review on recent advancements in therapy and management. AJRPSB 2015;3(1):12–26.

2. de Leeuw DC, van den Ancker W, Denkers F, de Menezes RX, Westers TM, Ossenkoppele GJ, et al. MicroRNA profiling can classify acute leukemias of ambiguous lineage as either acute myeloid leukemia or acute lymphoid leukemia. Clin Cancer Res 2013;19(8):2187–2196.

3. Sheikhpour R, Fazli R, Mehrabani S. Gene Identification from Microarray Data for Diagnosis of Acute Myeloid and



[ DOI: 10.18502/ijpho.v13i1.11629 ]

Lymphoblastic Leukemia Using a Sparse Gene Selection Method. IJPHO 2021;11(2):70–77.
4. Filippini T, Heck JE, Malagoli C, Giovane C del, Vinceti M. A review and meta-analysis of outdoor air pollution and risk of childhood leukemia. J Environ Sci Health - Part C 2015;33(1):36–66.
5. Masilamani V, Devanesan S, AlSalhi MS, AlQahtany FS, Farhat KH. Fluorescence spectral detection of acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML): A novel photodiagnosis strategy. Photodiagnosis Photodyn Ther 2020;29: 101634-101638.
6. Rasool M, Farooq S, Malik A, Shaukat A, Manan A, Asif M, et al. Assessment of circulating biochemical markers and antioxidative status in acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) patients. Saudi J Biol Sci 2015;22(1):106–111.
7. Hong JHH, Cho SBB. Gene boosting for cancer classification based on gene expression profiles. Pattern Recognit 2009;42(9):1761–1777.
8. Aziz R, Verma CK, Srivastava N. A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data. Genom Data 2016;8:4–15.
9. Elyasigomari V, Lee DA, Screen HRC, Shaheed MH. Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. J Biomed Inform 2017;67:11–20.
10. Apolloni J, Leguizamón G, Alba E. Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput 2016;38:922–932.
11. Chen Y, Zhang Z, Zheng J, Ma Y, Xue Y. Gene selection for tumor classification using neighborhood rough sets and entropy measures. J Biomed Inform 2017;67:59–68.
12. Nie F, Huang H, Cai X, Ding CH. Efficient and robust feature selection via joint l2,1-norms minimization. Advances in neural information processing systems. 2010. p. 1813–1821.
13. Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MAZ. A robust graph-based semi-supervised sparse feature selection method. Inf Sci 2020;531:13–30.
14. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 1999; 286(5439):531–537.
15. Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions. ICML 2003; 912–919.
16. Nie F, Xu D, Tsang IWH, Zhang C. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. IEEE Trans Image Process 2010;19(7):1921–1932.
17. Liaw A, Wiener M. RandomForest: Breiman and Cutler's random forests for classification and regression. R package version 2015;4:14-18.
18. Han J, Pei J, Kamber M. Data mining: concepts and techniques. domenica 2011;1-9.
19. Ma Z, Nie F, Yang Y, Uijlings JRR, Sebe N, Member S, et al. Discriminating joint feature analysis for multimedia data understanding. IEEE Trans Multimedia 2012;14(6):1662–1672.
20. Ma Z, Yang Y, Nie F, Uijlings J, Sebe N. Exploiting the entire feature space with sparsity for automatic image annotation. Proceedings of the 19th ACM international conference on Multimedia - MM 2011; 283-287.
21. Shi C, Ruan Q, An G. Sparse feature selection based on graph Laplacian for web image annotation. Image Vis Comput 2014; 32(3):189–201.
22. Arunkumar C, Ramakrishnan S. Attribute selection using fuzzy roughset



[ DOI: 10.18502/ijpho.v13i1.11629 ]

based customized similarity measure for lung cancer microarray gene expression data. FCIJ 2018;3(1):131–142.
23. Kumar A, Halder A. Ensemble-based active learning using fuzzy-rough approach for cancer sample classification. Eng Appl Artif Intell 2020;91: 103591-103594.
24. Halder A, Dey S, Kumar A. Active learning using fuzzy k-NN for cancer classification from microarray gene expression data. In: Advances in Communication and Computing. Springer 2015. 103–113.
25. Halder A, Kumar A. Active learning using rough fuzzy classifier for cancer prediction from microarray gene expression data. J Biomed Inform 2019; 92:103136-103140.
26. Sun L, Zhang XY, Qian YH, Xu JC, Zhang SG, Tian Y. Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. Appl Intell 2019;15;49(4):1245–1259.
27. Abdullah SK, Hasan SKR, Mollah AF. Acute Leukemia Subtype Prediction Using EODClassifier. In: Lecture Notes on Data Engineering and Communications Technologies. Springer Science and Business Media Deutschland GmbH 2022; 129–137.
28. Dey UK, Islam MS. Genetic expression analysis to detect type of leukemia using machine learning. In: 2019 1st international conference on advances in science, engineering and robotics technology (ICASERT) 2019; 1–6.
29. Maria IJ, Thirupathi D. Performance Analysis of Leukemic Gene Expression Profiles using Classification. In: Proceedings of the 5th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2021. Institute of Electrical and Electronics Engineers Inc.; 2021.1053–1058.
30. Li J, Wang Y, Jiang T, Xiao H, Song X. Grouped gene selection and multi-classification of acute leukemia via new regularized multinomial regression. Gene 2018; 667:18–24.
31. Wu Y, Ji R, Ge M, Shi S. Classification of Tumor Gene Expression Data Based on Manifold Learning and Gaussian Process. In: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) 2019; 1–5.



[ DOI: 10.18502/ijpho.v13i1.11629 ]