Original Article *Iran J Ped Hematol Oncol. 2024, Vol 14, No 2, 104-115*

In Silico Identification of Effective Genes for Acute Leukemia Classification Using a Spline Regression-based Framework

Maryam Yazdanparast[1], Razieh Sheikhpour[2*], Morteza Zangeneh Soroush[3], Fatemeh Ghanizadeh[4]

1. Department of Pediatrics, Shahid Sadoughi Hospital, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
2. Department of Computer Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran
3. Department of Biomedical Engineering, Science and Research branch, Islamic Azad University, Tehran, Iran
4. Hematology and Oncology Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

*Corresponding author: Dr Razieh Sheikhpour, Department of Computer Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran. Email: rsheikhpour@ardakan.ac.ir. ORCID ID: 0000-0002-3119-3349

Received: 05 January 2024 **Accepted:** 01 March 2024

Abstract

Background: Microarray technology enables the examination of gene expression in thousands of genes and can be highly effective in identifying various types of cancers, including leukemia. However, many genes in microarray data are redundant and lack useful information for cancer diagnosis. The main objective of this study is to identify relevant and effective genes in classification of leukemia microarray data using a spline regression-based method, taking into account the correlation between genes.

Materials and Methods: In this analytical study, leukemia microarray data are used to identify relevant genes in classification of leukemia into Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) using a spline regression-based gene selection method, called SRS^{3}FS based on $\ell_{2,p}$-norm ($0 < p \leq 1$). Subsequently, the support vector machine (SVM) algorithm is employed to classify leukemia data into AML and ALL.

Results: In this study, the classification results of SVM algorithm for 5, 10, 15, and 20 genes reveal that the SRS^{3}FS method, employing $\ell_{2,1/4}$-norm, $\ell_{2,1/2}$-norm and $\ell_{2,3/4}$-norm, exhibited the highest accuracy of 97.06% when identifying 10 genes for distinguishing between AML and ALL. Moreover, the leukemia data was classified into AML and ALL with an accuracy of 100%, using a gene identified by the SRS^{3}FS method based on $\ell_{2,3/4}$-norm and $\ell_{2,1}$-norm. The gene labeled as number 3252, annotated as GLUTATHIONE S-TRANSFERASE, MICROSOMAL, is recognized as the most important gene.

Conclusion: The experimental results on leukemia microarray data demonstrate that the spline regression-based gene selection method can effectively identify relevant genes in classification and prediction of leukemia.

Keywords: Acute lymphocytic leukemia, Acute myeloid leukemia, Gene expression, Sparse gene selection, Spline regression

Introduction

Cancer is a disease resulting from the uncontrolled division of cells and stands as one of the leading causes of death globally. Early-stage cancer diagnosis holds the potential for effective treatment (1). Leukemia is a prevalent and fatal type of cancer characterized by the uncontrolled proliferation and incomplete evolution of white blood cells and their precursors in the blood and bone marrow. Leukemia disrupts the regular growth and division control of white blood cells, diverting their growth from normal regulation (2). In this condition, an abnormal excess of blood cells, distinct from natural blood cells, is produced by the bone marrow, hindering the body's ability to combat diseases (3).

Leukemia affects the production of various types of blood cells, including red blood cells and platelets, which are produced by the bone marrow. Due to the severity and rapid progression of the disease, leukemia is categorized into acute and chronic forms. Depending on the involvement of white blood cell types, it is further classified into lymphocytic and myeloid types (4, 5). Acute Myeloid Leukemia

(AML) impacts myeloid cells or myeloblasts, displaying an acute course. Acute Lymphoblastic Leukemia (ALL) affects lymphocytic cells or lymphoblasts, also presenting an acute course (6). Successful treatment of AML requires distinguishing it from Acute Lymphoblastic Leukemia (ALL) (7).

One of the most crucial and precise techniques for diagnosing and predicting this disease involves analyzing individuals' DNA and their genetic information. A cutting-edge technology in gene expression data analysis is microarray technology, enabling simultaneous measurement of the expression of thousands of genes (6). In this technology, each known gene sequence of interest is printed on a glass or nylon array as a probe. These probes, labeled with fluorescent markers for mRNA from tissue or blood samples, are then hybridized onto an array (5, 7).

Microarray technology plays a crucial role in the diagnosis and prediction of various biological diseases, including cancer. It is employed to investigate gene expression and changes resulting from factors such as treatment, comparative genomic hybridization, identification of single nucleotide polymorphisms, and determination of the genomic content of living organisms, enabling their comparison (5, 7). In microarray experiments, gene expression data from thousands of genes are measured in different patients, generating high-dimensional data. One of the challenges in gene expression data using microarray technology is the limited number of samples from patients in contrast to the high dimensionality of genes, significantly impacting the accuracy of classification and often leading to its reduction (8-10).

Although a vast number of genes exist in microarray data, only a small fraction of them are necessary for cancer diagnosis and have a substantial impact on classification accuracy. Therefore, a crucial objective in the analysis of gene expression data is the identification and selection of relevant genes in the classification and diagnosis of cancer. Identifying these genes contributes to their use in classifiers and provides valuable information about the relationship between cancer and genes. This information can be instrumental in research across various types of cancer, including leukemia (11, 12). The selection of relevant genes in cancer is not possible without statistical analysis and gene selection methods (13).

Gene selection methods can be categorized into three main types: filter, wrapper, and embedded methods. In filter methods, genes are selected in a single preprocessing step without employing learning algorithms. These methods are widely used for high-dimensional microarray datasets due to their simplicity and speed. In wrapper methods, a subset of genes, maximizing the performance of a learning algorithm, is selected based on a learning algorithm. Despite the high accuracy of wrapper methods, they are not very efficient for high-dimensional microarray data due to their time-consuming nature. In embedded methods, gene selection and model learning are performed simultaneously using an objective function (14,15).

Filter methods are highly suitable for selecting efficient genes in microarray datasets with a large number of genes due to their speed (14). In classical filter gene selection methods, each gene is evaluated individually, neglecting the correlation between genes and their joint effects in the gene selection process. To address this challenge, sparse gene selection methods have been introduced, considering the correlation between genes in the gene selection process. In the methods that utilize sparse models, after obtaining the sparse transformation matrix for gene selection, the rank of each gene can be

calculated (13, 16). Gene selection methods based on sparse models, particularly $\ell_{2,1}$-norm sparse models, take into account the correlation between genes and have been widely used for sparse gene selection (17). The $\ell_{2,p}$-norm matrix ($0 < p \leq 1$) has been introduced as an extension of the $\ell_{2,1}$-norm, allowing for even sparser gene selection and potentially better performance compared to $\ell_{2,1}$-norm (18). The spline regression-based gene selection method, called SRS^{3}FS method, is a sparse gene selection approach that utilizes an $\ell_{2,p}$-norm-based regularization term and a spline regression matrix. Unlike methods using a Laplacian graph matrix, SRS^{3}FS better preserves the distribution and geometry structure of the data (16).

Most research conducted for the selection of relevant genes in classification of leukemia into AML and ALL cancer relies on classical gene selection methods. Given the aforementioned challenges, the objective of this study is to identify relevant genes in classifying leukemia into AML and ALL using an sparse spline regression-based gene selection method. Investigating the expression of these genes can potentially aid in the early diagnosis of various types of acute leukemias. After identifying relevant genes in leukemia, the support vector machine (SVM) algorithm is employed to classify leukemia microarray data into AML and ALL. The results obtained from the experiments indicate the effectiveness of the SRS^{3}FS based on $\ell_{2,3/4}$-norm in identifying relevant genes in classification of leukemia compared to other methods.

Materials and Methods

Data

The present study is an analytical study that utilizes gene expression data from patients with acute lymphoid and myeloid leukemia. The gene expression dataset for AML and ALL used in this study was provided by Golub et al. (7) and obtained from the bone marrow of leukemia patients. This dataset comprises expression data for 7129 genes from 72 leukemia patients, including 23 with AML and 49 with ALL, obtained through microarray technology. The dataset has been widely used by researchers and has been pre-divided into training and test sets. The training dataset is employed for gene selection and model construction, while the test dataset is utilized for model evaluation. The training consist of 38 leukemia patients, with 25 diagnosed with ALL and 13 with AML. The test dataset comprises 34 leukemia patients, including 24 with ALL and 10 with AML. Using the training dataset, effective genes in leukemia are identified through spline-regression-based gene selection method. Subsequently, the selected genes are used by the SVM algorithm for training a model. Finally, the model is evaluated on the test dataset to assess its performance on the selected genes.

Gene selection using spline regression-based framework

In this study, after collecting gene expression data for AML and ALL, data preparation is performed, including data normalization. Due to the large number of genes in leukemia microarray data, it is necessary to eliminate irrelevant and redundant genes and identify relevant genes for classification of leukemia into AML and ALL. Considering the high dimensionality of microarray data, filter gene selection methods are extensively employed to identify relevant genes in blood cancer. Classical filter gene selection methods evaluate the importance of genes individually (13). To overcome this challenge, sparse gene selection methods based on $\ell_{2,1}$-norm and $\ell_{2,p}$-norm ($0 < p \leq 1$) are introduced to jointly evaluate the genes, taking into account the correlation between them. In sparse gene

selection methods, an optimal transformation matrix $W \in R^{d\times c}$ is calculated to select genes, where each row of this matrix is used for weighting a specific gene.

In this study, a spline regression-based framework based on $\ell_{2,p}$-norm ($0 < p \leq 1$) called SRS^{3}FS is utilized to consider the correlation between genes during the gene selection process and select the most relevant genes in diagnosis of AML and ALL.

The SRS^{3}FS method is one of the sparse gene selection techniques that utilizes spline regression to preserve the geometric structure of the data. It incorporates a regularization term based on the $\ell_{2,p}$-norm to select relevant genes. This gene selection method is defined as Eq. (1) (16).

$$\underset{F,W,b}{\arg\min} Tr(W^T XGX^T W) + \mu\|X^T W + 1_n b^T - Y\|_F^2 + \lambda\|W\|_{2,p}^p, \quad p \in (0,1] \tag{1}$$

where, the term $Tr(W^T XGX^T W)$ preserves the geometric structure of the data, and G is the spline matrix. The term $\|X^T W + 1_n b^T - Y\|_F^2$ represents a loss function, where $b \in R^c$ is the bias term, and $1_n \in R^n$ is a column vector with all elements equal to one. The term $\|W\|_{2,p}^p$ is a regularization term based on $\ell_{2,p}$-norm ($0 < p \leq 1$), causing the rows of the transformation matrix to be sparse and appropriate for gene selection. μ and λ are the regularization parameters.

In Eq. (1), the predicted labels matrix $F = [f_1, f_2, \ldots, f_n]^T \in R^{n\times c}$ can be used instead of the actual labels matrix Y, where $f_i \in R^c \ (1 \leq i \leq n)$ is the predicted label for the sample x_i. In general framework of graph-Laplacian gene selection, the predicted labels should be close to the actual labels and smooth on the graph defined as Eq. (2).

$$\underset{F}{\arg\min} Tr(F^T LF) + Tr\left((F-Y)^T U(F-Y)\right) \tag{2}$$

where L is the graph Laplacian, $U \in R^{n\times n}$ is a diagonal matrix with infinite values on the main diagonal, referred to as the decision matrix, ensuring compatibility between the predicted labels by F and the actual labels Y.

Thus, by integration of Eq (1) and Eq. (2), the SRS^{3}FS framework is defined as follows:

$$\underset{F,W,b}{\arg\min} Tr(F^T GF) + Tr\left((F-Y)^T U(F-Y)\right) + \mu\|X^T W + 1_n b^T - F\|_F^2 + \lambda\|W\|_{2,p}^p, \quad p \in (0,1] \tag{3}$$

In Eq. (3), the local spline regression retains the local geometry structure of leukemia data, while the $\ell_{2,p}$-norm regularization ensures that this method selects the relevant genes.

Leukemia classification using support vector machine

SVM is a powerful classification algorithm widely utilized in the field of bioinformatics, particularly for the classification of leukemia data into AML and ALL. This method is well-suited for the task due to its ability to find an optimal hyperplane that maximally separates distinct classes in high-dimensional feature spaces. In the classification of leukemia, SVM operates by searching for a hyperplane that best discriminates between AML and ALL samples. The primary goal is to identify a hyperplane with the maximum margin, which is the distance between the hyperplane and the nearest data points of each class. By maximizing this margin, SVM aims to enhance the generalization ability of the classifier (19, 20).

Leukemia data often exhibit complex patterns and may not be linearly separable in their original feature space. SVM addresses this challenge by employing a kernel trick, which implicitly maps the input features into a higher-dimensional space. This transformation enables SVM to discover nonlinear relationships and

find an effective hyperplane for classification. The capability of SVM to handle nonlinear data patterns makes it particularly valuable in leukemia classification, where the relationships between gene expressions may not follow a linear trend. Furthermore, the ability of SVM to handle datasets with a relatively small number of samples, which is common in medical datasets, makes it well-suited for leukemia classification tasks. It avoids overfitting by focusing on the most relevant data points, known as support vectors, ensuring better generalization to unseen data (19, 20).

In summary, SVM is a potent tool for the classification of leukemia data into AML and ALL. Its capacity to handle nonlinear relationships and its adaptability to datasets with limited samples contribute to its effectiveness in extracting meaningful patterns from complex biological data.

Results

In this study, several experiments were conducted on microarray leukemia data to identify effective genes in leukemia classification using the SRS^{3}FS framework defined in Eq. (3). The regularization parameters in Eq. (3) were set to 1, and parameter p in $\ell_{2,p}$-norm-based regularization were fixed at {0.25, 0.5, 0.75, 1}. The performance of SRS^{3}FS was also compared to SFS (21) and Sr-SemiDFS (22), the sparse gene selection methods that utilize regularization based on $\ell_{2,1}$-norm.

Subsequently, using the SVM classification algorithm, modeling was carried out on the identified genes by the gene selection methods. After constructing the SVM model, the performance of the model in classifying leukemia data into AML and ALL was evaluated using the test dataset. The assessment was done based on the accuracy metric, aiming to determine the significance of the selected genes. The accuracy metric represents the percentage of leukemia data accurately classified by the SVM on the genes selected by the spline regression-based framework into AML and ALL.

The classification results of the SVM algorithm on 5, 10, 15, and 20 genes identified by SFS, Sr-SemiDFS and SRS^{3}FS methods are presented in Table I. The average results of these methods across different number of genes are also included.

As depicted in Table I, the gene selection method SRS^{3}FS based on $\ell_{2,3/4}$-norm and $\ell_{2,1}$-norm achieved higher accuracy in identifying 5 genes compared to other methods. The SRS^{3}FS method, utilizing $\ell_{2,1/4}$-norm, $\ell_{2,1/2}$-norm, and $\ell_{2,3/4}$-norm, demonstrated the highest accuracy when identifying 10 genes for distinguishing between AML and ALL. With the selection of 15 genes, SRS^{3}FS based on $\ell_{2,3/4}$ demonstrated higher accuracy compared to other methods. Moreover, using 20 selected genes, SRS^{3}FS based on $\ell_{2,3/4}$ exhibited higher accuracy than other methods. The average accuracy of the SVM classification algorithm, with a varying number of genes selected by different gene selection methods, indicated that SRS^{3}FS based on $\ell_{2,3/4}$ achieved the highest accuracy compared to other methods. The SRS^{3}FS method based on $\ell_{2,3/4}$ showed the highest accuracy with 5, 10, 15, and 20 genes selected compared to other methods.

Since all gene selection methods achieved the highest accuracy with the identification of 10 genes, subsequent experiments will investigate the performance of various gene selection methods from 1 to 10 genes. Figure 1 illustrates the classification accuracy of SVM algorithm utilizing 1 to 10 selected genes, employing SFS, Sr-SemiDFS and SRS^{3}FS methods.

The results obtained in Figure 1 indicated that the SRS^{3}FS gene selection method

based on $\ell_{2,3/4}$-norm and $\ell_{2,1}$-norm, achieved 100% accuracy with just one gene, effectively identifying all instances of ALL and AML. Furthermore, Figure 1 illustrated that the performance of the SRS^{3}FS method based on both $\ell_{2,3/4}$-norm and $\ell_{2,1}$-norm, remained superior with varying number of genes compared to other methods. Table II presents the average classification accuracy of the SVM algorithm using 1 to 10 selected genes for leukemia, employing different gene selection methods. Results in Table II demonstrated that the average classification accuracy of the SVM algorithm, when considering 1 to 10 selected genes for leukemia utilizing the SRS^{3}FS gene selection method based on $\ell_{2,3/4}$-norm, surpassed that of other methods. Additionally, the SRS^{3}FS gene selection method using $\ell_{2,1}$-norm was ranked second. The findings from Tables 1 and 2, and Figure 1 highlighted the efficacy of the SRS^{3}FS gene selection method based on $\ell_{2,3/4}$-norm in identifying relevant genes for leukemia classification. These results indicated that the SRS^{3}FS gene selection method based on $\ell_{2,3/4}$-norm can accurately identify ALL and AML with a minimal number of genes. In the subsequent experiments, the 10 genes identified by the SRS^{3}FS gene selection method based on $\ell_{2,3/4}$-norm and $\ell_{2,1}$-norm are further examined, and their respective rankings are presented in Tables III and IV. As seen from Tables III and IV, the gene with number 3252 has been identified as the most important gene by the SRS^{3}FS gene selection method based on $\ell_{2,3/4}$-norm and $\ell_{2,1}$-norm. Furthermore, as indicated in the results from Figure 1, the SVM algorithm, using this gene, can accurately detect ALL and AML with 100% accuracy. Among the identified genes, six genes highlighted in Table V were jointly selected by the SRS^{3}FS gene selection method based on $\ell_{2,3/4}$-norm and $\ell_{2,1}$-norm.

Table I: Classification accuracy of the SVM algorithm on selected genes in leukemia data by sparse gene selection methods SFS, Sr-SemiDFS and SRS^{3}FS

Gene selection method	Accuracy (5 genes)	Accuracy (10 genes)	Accuracy (15 genes)	Accuracy (20 genes)	Average accuracy
SFS	44.12	94.12	91.18	41.18	67.65
Sr-SemiDFS	35.29	91.18	41.18	73.53	60.30
SRS^{3}FS ($\ell_{2,1/4}$)	38.24	**97.06**	47.06	44.12	56.62
SRS^{3}FS ($\ell_{2,1/2}$)	38.24	**97.06**	47.06	44.12	56.62
SRS^{3}FS ($\ell_{2,3/4}$)	**88.24**	**97.06**	**97.06**	**97.06**	**94.86**
SRS^{3}FS ($\ell_{2,1}$)	**88.24**	94.12	38.24	41.18	65.45

Table II: Average classification accuracy of the SVM algorithm using 1 to 10 selected genes for leukemia, employing SFS, Sr-SemiDFS and SRS^{3}FS methods

Gene selection method	Average accuracy
SFS	53.84
Sr-SemiDFS	52.29
SRS^{3}FS ($\ell_{2,1/4}$)	56.18
SRS^{3}FS ($\ell_{2,1/2}$)	56.18
SRS^{3}FS ($\ell_{2,3/4}$)	**93.83**
SRS^{3}FS ($\ell_{2,1}$)	92.06

Table III: 10 genes selected by the SRS^{3}FS gene selection method based on $\ell_{2,3/4}$-norm

Gene Number	Accession	Description
3252	U46499_at	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
2043	M57710_at	LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)
3847	U82759_at	GB DEF = Homeodomain protein HoxA9 mRNA
4443	X66533_at	GUANYLATE CYCLASE SOLUBLE, BETA-1 CHAIN
4279	X55668_at	PRTN3 Proteinase 3 (serine proteinase, neutrophil, Wegener granulomatosis autoantigen)
1779	M19507_at	MPO Myeloperoxidase
2363	M93056_at	LEUKOCYTE ELASTASE INHIBITOR
4653	X80907_at	GB DEF = P85 beta subunit of phosphatidyl-inositol-3-kinase
6215	M19508_xpt3_s_at	MPO from Human myeloperoxidase gene, exons 1-4./ntype=DNA /annot=exon
5954	Y00339_s_at	CA2 Carbonic anhydrase II

Table IV: 10 genes selected by the SRS^{3}FS gene selection method based on $\ell_{2,1}$-norm

Gene Number	Accession	Description
3252	U46499_at	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
2043	M57710_at	LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)
2288	M84526_at	DF D component of complement (adipsin)
5039	Y12670_at	LEPR Leptin receptor
4443	X66533_at	GUANYLATE CYCLASE SOLUBLE, BETA-1 CHAIN
4052	X04085_rna1_at	Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)
3847	U82759_at	GB DEF = Homeodomain protein HoxA9 mRNA
1779	M19507_at	MPO Myeloperoxidase
2181	M68891_at	GATA2 GATA-binding protein 2
4653	X80907_at	GB DEF = P85 beta subunit of phosphatidyl-inositol-3-kinase

Table V: Genes selected jointly by the SRS^{3}FS gene selection method based on $\ell_{2,3/4}$-norm and $\ell_{2,1}$-norm

Gene Number	Accession	Description
3252	U46499_at	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
2043	M57710_at	LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)
3847	U82759_at	GB DEF = Homeodomain protein HoxA9 mRNA
4443	X66533_at	GUANYLATE CYCLASE SOLUBLE, BETA-1 CHAIN
1779	M19507_at	MPO Myeloperoxidase
4653	X80907_at	GB DEF = P85 beta subunit of phosphatidyl-inositol-3-kinase

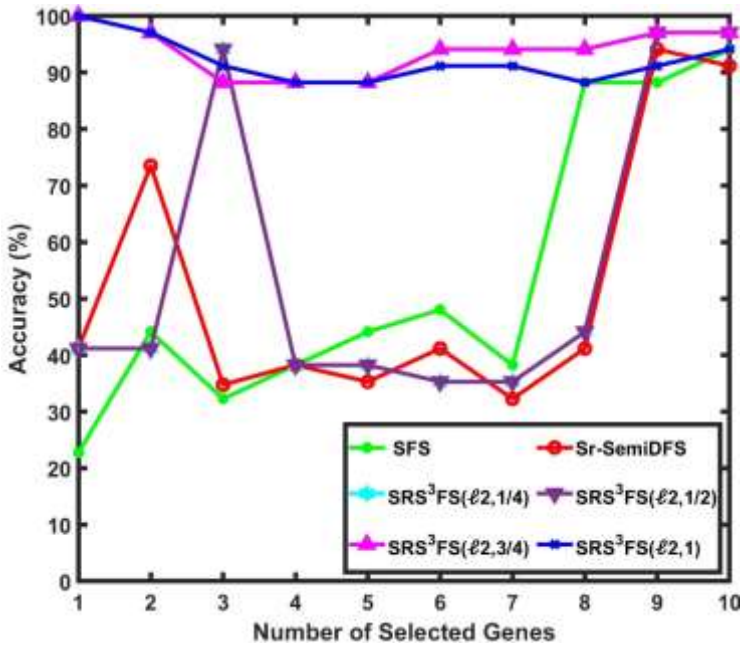

Figure 1. Classification accuracy of the SVM algorithm using 1 to 10 selected genes for leukemia, employing SFS, Sr-SemiDFS, and SRS^{3}FS methods

Discussion

One of the significant technologies that has enabled the simultaneous examination of thousands of genes for cancer detection is microarray technology. A major challenge in analyzing microarray data is the high dimensionality and the limited number of patients, which complicates the classification algorithms, reduces their generalization ability, and increases computational costs. Therefore, the identification and selection of effective genes are crucial steps in predicting and diagnosing cancer (23, 24).

Most studies on the identification of effective genes in classification of leukemia have utilized classical filter gene selection methods. Classical filter gene selection methods do not consider the correlation between genes when identifying them and select genes one by one. Hence, in this study, the identification of important and relevant genes in classification of leukemia into AML and ALL was carried out using the SRS^{3}FS, a spline regression-based gene selection method based on $\ell_{2,p}$-norm ($0 < p \leq 1$), which consider the correlation between genes. After identifying and selecting relevant and effective genes, the SVM algorithm was employed to classify and predict microarray data into two categories: ALL and AML. The performance was evaluated using accuracy metrics. The results of the experiments on leukemia microarray data showed that the SRS^{3}FS method based on $\ell_{2,3/4}$-norm and $\ell_{2,1}$-norm identified the most relevant gene in leukemia, and the SVM algorithm classified all samples correctly using only one gene, achieving 100% accuracy in distinguishing ALL and AML.

In a study by Li et al., genetic algorithm and k-nearest neighbors classification achieved an accuracy of 86.4% on gene expression data for leukemia (25).

Kumar and Halder presented an ensemble learning method using a fuzzy algorithm

for classifying microarray data, achieving an accuracy of 97.72% (26). They also employed the ALFKNN method, achieving an accuracy of 93.62%. ALFKNN selected the most uncertain samples and classified the test set using fuzzy k-nearest neighbors (27). Furthermore, Kumar and Halder classified leukemia with an accuracy of 98.88% using the ALRFC method (28).

Arunkumar and Ramakrishnan proposed a similarity measure for gene selection using a fast fuzzy algorithm. In the first step, they utilized an entropy-based method for dimensionality reduction of gene expression data, and in the second step, they applied the fast fuzzy algorithm, which defines a similarity measure, to identify the minimum number of effective genes in leukemia. This fuzzy method achieved an accuracy of 90.28% on the initial gene expression data and, after dimensionality reduction in the first step, attained an accuracy of 97.22% on the reduced data (29).

Wang et al. achieved accuracies of 72.64% and 87.5% in the classification of AML and ALL patients using the k-nearest neighbors and single NF methods, respectively (30).

Cai et al. employed the I-RELEF-NB method and obtained an accuracy of 91.67% for classifying leukemia into AML and ALL. They achieved an accuracy of 92.86% by applying the I-RELIEF-LDA method to these data, and with the RELIEF-KNN method, they achieved an accuracy of 94.44% in the classification of AML and ALL (31).

Zhang et al. used BMSF-NB and Gene SrF-NB methods, achieving accuracies of 96.5% and 94.58%, respectively, in detecting AML and ALL (32).

Dey and Islam predicted ALL and AML using three machine learning algorithms. Initially, they employed Principal Component Analysis (PCA) to reduce the dimensions of gene expression data for leukemia. Subsequently, utilizing artificial neural networks, random forest classifier, and XGBoost algorithm, they achieved accuracies of 92.3%, 80.8%, and 92.3%, respectively, in predicting different types of acute leukemia (33).

Sheikhpour et al. using various classification algorithms and the sparse gene selection method based on $\ell_{2,1}$-norm, achieved classification accuracies of 94.12%, 100%, 91.18%, and 100% for SVM, k-nearest neighbors, gaussian kernel density estimation, and linear discriminant analysis, respectively in classification of leukemia into AML and ALL (13).

Cho and Won employed k-nearest neighbors and SVM with a gaussian kernel on microarray data, achieving accuracies of 91.6% and 94.4%, respectively (34).

Nguyen and Rocke achieved accuracy of 94.4% in detecting leukemia using the logistic regression method and accuracy of 95.4% with second order discriminant analysis (35).

Mehrabani et al. employed leukemia gene expression data to assess the significance of each gene through GS^{3}FS, an $\ell_{2,p}$-norm sparsity-based gene selection method. They identified effective genes for classifying leukemia into AML and ALL using Random Forest (RF) and SVM classifiers. The RF classifier accurately classified all AML and ALL samples, achieving accuracy of 100% with 10 genes selected by the GS^{3}FS method based on $\ell_{2,1/2}$-norm and $\ell_{2,1}$-norm. The optimal performance of the RF algorithm in classifying AML and ALL reached 100% accuracy. This achievement was realized by utilizing 15 genes selected through the GS^{3}FS method based on $\ell_{2,1/4}$-norm, $\ell_{2,1/2}$-norm, and $\ell_{2,1}$-norm (36).

Chen and Lin using a backpropagation neural network on leukemia gene expression data, achieved an accuracy of 95.83%. With the MTSVSL method, they

reached an accuracy of 96.67% on leukemia data (37).

Conclusion

The findings of this study highlight the effectiveness of the spline regression-based gene selection method when paired with the SVM classification algorithm, providing accurate predictions for various types of leukemia. This gene selection method identifies the most relevant genes for classification and prediction of leukemia. Further analysis of these selected genes present a promising avenue for improving the predictive capabilities in leukemia diagnosis.

Ethical Considerations

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

Acknowledgements

None

Funding

None

Author contributions

Maryam Yazdanparast: Conceptualization; Investigation; Methodology; Validation; Writing - original draft.
Razieh Sheikhpour: Visualization; Formal analysis; Investigation; Methodology; Validation; Writing - original draft; Writing - review & editing.
Fatemeh Ghanizadeh: Investigation; Validation; Editing

Conflict of interest

The authors declare no conflict of interest.

References

1. Sheikhpour R, Hekmat Moghadam H. The effect of estrogen on p53 protein in T47D breast cancer cell line. Razi J Med Sci 2015; 22(133): 51-58.
2. Torkaman A, Charkari NM, Aghaeipour M. An approach for leukemia classification based on cooperative game theory. Anal Cell Pathol(Amst) 2011; 34(5): 235-246.
3. Zand AM, Imani S, Saadati M, Borna H, Ziaei R, Honari H. Effect of age, gender and blood group on blood cancer types. Kowsar Med J 2010 15(2): 111-114.
4. Toloie Ashlaqi A, Mohsen Taheri S.Designing an expert system for suggesting the bloodcancer treatment J Health Admin 2010; 13(40): 41-50.
5. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet 1999; 21(1 Suppl): 10-14.
6. Heidari M, Hajigholami A. Acute lymphocytic leukemia with severe eosinophilia (a case report). J Shahrekord Univ Med Sci 2013; 15(5): 111-115.
7. 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(15): 530-538.
8. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Distributed feature selection: An application to microarray data classification. Appl Soft Comput 2015; 30: 136-150.
9. Martinez E, Alvarez MM, Trevino V. Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm. Comput Biol Chem 2010; 34(4): 244-250.
10. Chu W, Ghahramani Z, Falciani F, Wild DL. Biomarker discovery in microarray gene expression data with Gaussian processes. Bioinformatics 2005; 21(16): 3385-3393.
11. 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.
12. 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.
13. Sheikhpour R, Fazli R, Mehrabani S. Gene Identification from Microarray Data for Diagnosis of Acute Myeloid and Lymphoblastic Leukemia Using a Sparse Gene Selection Method. Iran J Pediatr Hematol Oncol 2021;11(2): 70-77.
14. Nie F, Huang H, Cai X, Ding C. Efficient and robust feature selection via joint ℓ2,1-norms minimization. Adv Neural Inf Process 2010; 23: 1813–1821.
15. Sheikhpour R, Sarram MA. Diagnosis of diabetes using an intelligent approach based on bi-level dimensionality reduction and classification algorithms, Iran J Diabetes Obes 2014; 10; 6(2): 74-84.
16. Sheikhpour R. A local spline regression-based framework for semi-supervised sparse feature selection. Knowledge-Based Syst 2023; 262: 110265-110268.
17. Ma Z, Nie F, Yang Y, Uijlings JR, Sebe N, Hauptmann AG. Discriminating joint feature analysis for multimedia data understanding. IEEE Trans Multimedia 2012; 14 (6): 1662-1672.
18. Wang L, Chen S. l2, p-Matrix Norm and Its Application in Feature Selection. arXiv preprint arXiv 2013:1303.3987.
19. Karim A, Azhari A, Shahroz M, Belhaouri SB, Mustofa K. LDSVM: Leukemia cancer classification using machine learning. Comput Mater Contin 2021; 71 (2): 3887-3903.
20. Hakim MA, Adiwijaya A, Astuti W. Comparative analysis of Relief F-SVM and CFS-SVM for microarray data classification. Int J Electr Comput Eng 2021; 11 (4): 3393.
21. Li X, Zhang Y, Zhang R. Semisupervised feature selection via generalized uncorrelated constraint and manifold embedding. IEEE Trans Neural Networks Learn Syst 2022; 33 (9): 5070-5079.
22. Fan M, Zhang X, Hu J, Gu N, Tao D. Adaptive data structure regularized multiclass discriminative feature selection. IEEE Trans Neural Networks Learn Syst 2022 21; 33 (10): 5859-5872.
23. Alshamlan HM, Badr GH, Alohali YA. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Comput. Biol. Chem 2015; 56: 49-60.
24. Tabakhi S, Najafi A, Ranjbar R, Moradi P. Gene selection for microarray data classification using a novel ant colony optimization. Neurocomputing 2015; 168: 1024-36.
25. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001; 17(12): 1131-1142.
26. Kumar A, Halder A. Ensemble-based active learning using fuzzy-rough approach for cancer sample classification. Eng Appl Artif Intell 2020; 91: 103591-10395.
27. 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 2015. Springer, New Delhi 2015;103-113.
28. Halder A, Kumar A. Active learning using rough fuzzy classifier for cancer prediction from microarray gene expression data. J Biomed Inf 2019; 92: 103136-103140.
29. Arunkumar C, Ramakrishnan S. Attribute selection using fuzzy rough set

based customized similarity measure for lung cancer microarray gene expression data. Future Comput Inf J 2018; 3(1): 131-142.
30. Wang Z, Palade V, Xu Y. Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis. In2006 international symposium on evolving fuzzy systems IEEE 2006; 241-246.
31. Cai H, Ruan P, Ng M, Akutsu T. Feature weight estimation for gene selection: a local hyperlinear learning approach. BMC Bioinf 2014; 15 (1): 1-13.
32. Zhang H, Wang H, Dai Z, Chen MS, Yuan Z. Improving accuracy for cancer classification with a new algorithm for genes selection. BMC Bioinf 2012; 13(1):1-20.
33. Dey UK, Islam MS. Genetic expression analysis to detect type of leukemia using machine learning. In2019 1st international conference on advances in science, engineering and robotics technology (ICASERT) IEEE 2019; 1-6.
34. Cho SB, Won HH. Machine learning in DNA microarray analysis for cancer classification. InProceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003; 19: 189-198.
35. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics2002; 18 (1): 39-50.
36. Mehrabani S, Soroush MZ, Kheiri N, Sheikhpour R, Bahrami M. Prediction of blood cancer using leukemia gene expression data and sparsity-based gene selection methods. Iran J Pediatr Hematol Oncol 2023; 13 (1): 13-21.
37. Chen AH, Lin EJ. The prediction of cancer classification using a novel multi-task support vector sample learning technique. AISS: Adv Inform Sci Serv Sci 2011; 3 (3): 92-99.