Iran J Ped Hematol Oncol. 2026, Vol. 16, No. 3, 934-947

Original Article

Adaptive Structure Learning for Leukemia Biomarker Discovery from Gene Expression Data

Razieh Sheikhpour [1*] PhD, Morteza Zangeneh Soroush[2] PhD, Azam Sadat Hashemi [3] MD

[1*] Department of Computer Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran
[2] Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[3] Hematology and Oncology Research Center, Noncommunicable Diseases Research Institute, 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: May 14, 2026
Accepted: June 10, 2026

Abstract

Background: Leukemia classification based on gene expression data is a challenging problem due to the high dimensionality of microarray datasets and the limited number of patient samples. Identifying a small subset of informative genes (biomarkers) that can accurately distinguish between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) is essential for improving diagnostic accuracy and supporting precision medicine approaches.

Materials and Methods: In this methodological study, an adaptive structure learning framework is proposed for biomarker discovery using the Golub Leukemia Dataset. The proposed method jointly integrates adaptive similarity structure learning, sparse feature selection, and sample reweighting into a unified optimization model. This framework learns both the intrinsic geometric structure of the data and the most discriminative gene subset simultaneously. The selected genes were evaluated using two classifiers, including k-nearest neighbor (KNN) and support vector machine (SVM).

Results: Experimental results show that the proposed method achieves 97.06% classification accuracy using KNN with 8 selected genes and 100% accuracy using SVM with only 7 genes. Comparative analysis with existing feature selection methods demonstrates that the proposed approach achieves superior or competitive performance while using a significantly smaller number of genes.

Conclusion: The proposed framework effectively identifies compact and highly discriminative biomarker sets for leukemia classification. By jointly modeling sample relationships and gene relevance, the method improves classification performance while reducing feature dimensionality. The results suggest that the proposed framework has potential applications in clinical decision-support systems and other high-dimensional biomedical classification problems.

Keywords: Biomarker, Data, Discovery, Gene expression

[DOI: 10.18502/ijpho.v16i3.21751]

Introduction

Leukemia is a group of malignant hematological disorders characterized by the uncontrolled proliferation and accumulation of abnormal leukocytes in the bone marrow, peripheral blood, and occasionally other tissues. These malignant cells disrupt normal hematopoiesis, leading to anemia, immunosuppression, and bleeding disorders. Clinically, leukemia is broadly classified into four main subtypes based on the rate of progression (acute or chronic) and the lineage of affected cells (lymphoid or myeloid): acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL), and chronic myeloid leukemia (CML) (1-3). Among these, ALL and AML are acute forms that progress rapidly and require immediate diagnosis and treatment, making their accurate differentiation particularly critical in clinical practice (1).

ALL originates from the malignant transformation of immature lymphoid progenitor cells, primarily affecting children but also occurring in adults. In contrast, AML arises from myeloid lineage precursors and is more common in older adults. Although both subtypes share overlapping clinical symptoms such as fatigue, fever, and bone marrow failure, their biological mechanisms and treatment strategies differ significantly (3,4). Therefore, precise and early classification between ALL and AML is essential for selecting appropriate chemotherapy regimens and improving patient survival rates.

Traditional diagnostic approaches for leukemia classification include morphological examination of peripheral blood and bone marrow smears, immunophenotyping using flow cytometry, cytogenetic analysis, and molecular testing. Although these methods are widely used in clinical practice, they are often time-consuming, require specialized expertise, and may suffer from inter-observer variability and limited sensitivity in early-stage disease detection (4).

With the advent of high-throughput genomic technologies such as DNA microarrays and next-generation sequencing (NGS), it has become possible to measure the expression levels of thousands of genes simultaneously (5-7). These technologies have significantly advanced cancer research by enabling the discovery of molecular biomarkers that reflect disease subtypes at the transcriptomic level (8-9). In particular, gene expression profiling has been widely used for leukemia classification, where patterns of gene activity can distinguish between ALL and AML with high accuracy (7,10). However, such datasets are typically characterized by extremely high dimensionality and limited sample sizes, leading to the well-known “small n, large p” problem. This imbalance poses significant challenges for traditional statistical and machine learning models, often resulting in overfitting and poor generalization (7).

To address these challenges, artificial intelligence (AI) and machine learning (ML) methods have been increasingly applied in leukemia research for diagnosis, risk stratification, and biomarker identification (11). Recent studies have demonstrated that ML models such as support vector machines (SVM), random forests, gradient boosting, and deep learning can effectively analyze high-dimensional omics data and identify clinically relevant biomarkers (7,10-13). For example, Cheng et al. (10) reported that AI-based models achieved high diagnostic performance in hematological malignancies, particularly in AML.

In addition, multiple machine learning-based studies have shown that feature selection methods can significantly improve biomarker discovery by reducing dimensionality and identifying robust gene signatures associated with AML and ALL prognosis and diagnosis (10, 13,14). These methods generally include filter-based techniques, wrapper methods, embedded feature selection models, and hybrid approaches. Filter methods evaluate genes based on statistical measures such as t-tests or information gain, while wrapper methods rely on classifier performance to select optimal subsets of features (15,16). Embedded methods, such as LASSO and support vector machine recursive feature elimination (SVM-RFE), integrate feature selection directly into the learning process (10). Despite these advances, many existing approaches treat samples independently and fail to capture the intrinsic

[DOI: 10.18502/ijpho.v16i3.21751]

geometric structure of gene expression data, which may contain important biological relationships among patients.

Recent advances in graph-based learning have demonstrated that incorporating structural relationships between samples can significantly improve classification performance in biomedical data analysis. Graph-based methods construct similarity networks where nodes represent samples and edges capture pairwise relationships. However, most conventional approaches rely on fixed or pre-defined similarity graphs, which may not accurately reflect the complex and noisy structure of gene expression data (17,18). This limitation has motivated the development of adaptive structure learning techniques that dynamically update sample relationships during the learning process.

In this study, an adaptive structure learning framework is proposed for leukemia biomarker discovery using gene expression data from the well-known Golub dataset. The framework integrates adaptive similarity graph learning, sparse feature selection, and sample reweighting into a unified optimization model. By jointly learning the intrinsic geometric structure of samples and identifying the most informative genes, the proposed approach aims to improve classification accuracy while selecting a compact and biologically meaningful subset of biomarkers for distinguishing between ALL and AML.

Material and Methods

This section describes the leukemia gene expression dataset, preprocessing procedure, and the proposed adaptive structure learning framework for leukemia biomarker discovery. The overall workflow of the proposed method is illustrated in Figure 1. The framework consists of several major stages, including leukemia gene expression data preprocessing, adaptive structure learning, sparse biomarker selection and leukemia subtype classification using machine learning algorithms. Initially, the microarray gene expression data are normalized and filtered to remove noisy genes. Subsequently, an adaptive similarity matrix is learned to dynamically model the intrinsic geometric relationships among leukemia samples. Simultaneously, sparse feature selection is performed to identify the most discriminative genes associated with AML and ALL classification. Finally, the selected biomarkers are evaluated using machine learning classifiers, including KNN and SVM, to assess their discriminative capability in leukemia subtype prediction.

Leukemia Gene Expression Data

In this methodological study, the publicly available Golub Leukemia Dataset was used to evaluate the proposed biomarker discovery framework. This benchmark dataset, originally introduced by Todd Golub and colleagues, contains gene expression profiles derived from bone marrow samples of patients diagnosed with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) (19).

The dataset consists of 72 samples, including 47 ALL and 25 AML cases, with expression levels measured across 7129 genes. Following the original experimental protocol, the data were partitioned into a training set (38 samples: 25 ALL and 13 AML) and a test set (34 samples: 22 ALL and 12 AML).

Data Preprocessing

Gene expression values were normalized using z-score normalization to reduce scale variability across genes. Genes with near-constant expression across samples were removed to reduce noise and improve computational efficiency. Let X∈R^(d×n) denote the gene expression matrix, where dand nrepresent the number of genes and samples, respectively. The corresponding label matrix is defined as Y∈R^(n×c) where c is the number of classes.

Adaptive Structure Learning

Conventional graph-based feature selection methods rely on a fixed similarity graph constructed in the original feature space (20,21). However, such predefined structures may not adequately capture the intrinsic geometry of noisy and high-dimensional leukemia data. To overcome this limitation, an adaptive similarity learning strategy is introduced, which dynamically updates

[DOI: 10.18502/ijpho.v16i3.21751]

the sample relationships during the learning process. Let S∈R^(n×n) denote the similarity matrix, where S_ij measures the relationship between samples x_iand x_j. The adaptive graph is learned by solving

$$\min_{S} \sum_{i,j} \| W^T x_i - W^T x_j \|^2 S_{ij} + \alpha \| S \|_F^2 \quad \text{s.t. } S_{ij} \geq 0, \sum_j S_{ij} = 1 \tag{1}$$

This formulation preserves local geometric structure while avoiding degenerate solutions through Frobenius norm regularization.

Robust Sparse Biomarker Discovery

To jointly perform biomarker selection and robust classification, adaptive structure learning is integrated with sparse feature selection and sample reweighting. The final optimization problem is defined as:

$$\min_{W,S} \| V(X^T W - Y) \|_F^2 + \lambda \sum_{i,j} \| W^T x_i - W^T x_j \|^2 S_{ij} + \beta \| W \|_{2,1} + \alpha \| S \|_F^2 \quad \text{s.t } S_{ij} \geq 0, \sum_j S_{ij} = 1, v_i \geq 0, \sum_i v_i = n \tag{2}$$

where $W \in \mathbb{R}^{d \times c}$ is the projection matrix used for biomarker selection, and $V = \mathrm{diag}(v_1, \ldots, v_n)$ is a diagonal sample-weighting matrix. The sample weights are adaptively updated according to:

$$v_i = \frac{1}{2 \| W^T x_i - y_i \| + \epsilon} \tag{3}$$

where ϵ is a small positive constant to avoid numerical instability.

The weighted classification term reduces the influence of noisy or outlier leukemia samples during learning. Samples associated with larger prediction errors automatically receive smaller weights, improving the robustness of biomarker selection. The second term preserves the intrinsic geometric structure of leukemia samples through adaptive similarity learning. The third term imposes row sparsity on the projection matrix W, enabling the selection of a compact subset of informative genes. Genes corresponding to rows with larger l_2-norm values are considered more important biomarkers.

The proposed optimization problem was solved using an alternating optimization strategy. Specifically, the projection matrix Wand adaptive similarity matrix S were iteratively updated until convergence. First, the similarity matrix was updated while fixing the projection matrix. Subsequently, the projection matrix was recalculated using the updated similarity structure and sample weights. This iterative process continued until the objective function stabilized.

Biomarker Ranking and Evaluation

After convergence, genes were ranked based on the ℓ2-norm of rows of the projection matrix W. Genes with larger values were selected as candidate biomarkers. To evaluate classification performance, two standard classifiers KNN and SVM were used. Model performance was evaluated using several standard metrics. Accuracy measures the overall proportion of correctly classified leukemia samples (AML and ALL), reflecting the general effectiveness of the model. Sensitivity (recall) quantifies the ability of the model to correctly identify AML patients, which is particularly important for detecting true positive cancer cases. Specificity evaluates how well the model correctly identifies ALL samples as negative cases, indicating its ability to avoid false alarms. Precision reflects the reliability of AML predictions by measuring how many predicted AML cases are truly AML. The F1-score provides a balanced measure between precision and recall, especially useful when the class distribution is not perfectly balanced. Finally, the area under the receiver operating characteristic curve (AUC-ROC) evaluates the overall discriminative ability of the model to distinguish between AML and ALL across all possible classification thresholds, providing a threshold-independent performance measure.

[DOI: 10.18502/ijpho.v16i3.21751]

KNN Classifier

KNN algorithm is a simple yet effective non-parametric classification method that is widely used in biomedical pattern recognition tasks. In this study, KNN is employed as one of the classifiers to evaluate the discriminative power of the selected leukemia biomarkers for distinguishing between ALL and AML. The method classifies each test sample based on the majority label of its K nearest neighbors in the feature space, where similarity is typically measured using a distance metric such as Euclidean distance. Unlike model-based approaches, KNN does not assume any prior distribution of the data and makes predictions solely based on local neighborhood information. This property makes it particularly suitable for high-dimensional gene expression data after feature selection, where informative genes enhance class separability. However, the performance of KNN is sensitive to the choice of K and the quality of the feature representation, which highlights the importance of effective biomarker selection in improving classification accuracy (22).

SVM Classifier

SVM is a widely used and powerful supervised learning algorithm for binary classification tasks, including biomedical applications such as leukemia subtype discrimination. In this study, SVM is employed to classify samples into ALL and AML based on the selected gene expression features. The main objective of SVM is to construct an optimal separating hyperplane that maximizes the margin between the two classes, thereby improving generalization performance on unseen data. For high-dimensional and complex gene expression data, where linear separability is often not guaranteed, SVM utilizes kernel functions to implicitly map the input data into a higher-dimensional feature space. This transformation enables the algorithm to capture nonlinear relationships and achieve better class separation between AML and ALL samples. Due to its robustness in handling high-dimensional datasets and its strong theoretical foundation, SVM is particularly suitable for microarray-based leukemia classification problems (22).

Results

To evaluate the effectiveness of the proposed adaptive structure-preserving biomarker discovery method, extensive experiments were conducted on the Golub leukemia gene expression dataset. The proposed method was compared with six state-of-the-art gene selection methods, including GS3FS (10), RS3FS (23), SFS (24), Sr-SemiDFS (25), ASLCGLFS (26), AGLSOFS (27) and SFS-SLL (28), using two well-known classifiers KNN and SVM.

The proposed framework was evaluated using a fixed training and test split as described in the dataset section. All experiments were conducted using identical parameter settings across all compared methods to ensure a fair evaluation. The regularization parameters λ, α, and β were tuned empirically based on cross-validation performance on the training set. Specifically, a grid search strategy was employed to select the optimal parameter values that maximize classification performance while maintaining model sparsity and stability. In addition, cross-validation was used to determine the optimal number of selected genes. The feature selection process was repeated for different numbers of top-ranked genes, and the subset yielding the best average classification performance across validation folds was selected as the final biomarker set.

For classification, the KNN classifier was implemented with K=3, and the SVM classifier was used with a radial basis function (RBF) kernel. The optimization process was iteratively performed until convergence based on the stability of the objective function.

Performance Evaluation Using KNN Classifier

Table 1 presents the classification results using the KNN classifier. The proposed method achieved the highest classification accuracy of 97.06% using only 8 selected genes. In addition, the proposed framework obtained 100% sensitivity, indicating that all AML samples were correctly identified. The method also achieved high specificity (95.83%), precision (90.91%), and F1-score (95.24%), demonstrating a balanced and robust

[DOI: 10.18502/ijpho.v16i3.21751]

classification performance. The proposed method further achieved the best ROC performance with an AUC-ROC value of 1.00, indicating superior discrimination capability compared with the competing approaches.

Compared with the previously developed GS3FS method, the proposed framework improved sensitivity, F1-score, and ROC performance while using fewer informative genes. Although Sr-SemiDFS achieved a sensitivity of 100%, its overall accuracy was only 47.06%, indicating a severe imbalance in classification performance. The proposed method achieved perfect sensitivity while maintaining high specificity (95.83%), demonstrating a well-balanced classification performance. These findings suggest that adaptive structure learning and robust sample weighting improve the identification of discriminative leukemia-related biomarkers.

Performance Evaluation Using SVM Classifier

To further evaluate the robustness of the selected biomarkers, classification experiments were also performed using the SVM classifier as shown in Table 2. The results presented in Table 2 demonstrate that the proposed method achieved the best overall classification performance among all compared approaches. Specifically, the proposed framework achieved 100% accuracy, sensitivity, specificity, precision, and F1-score using only 7 selected genes. Moreover, the proposed method achieved the highest ROC performance with an AUC-ROC value of 1.00 indicating excellent discrimination between AML and ALL samples. Although GS3FS and RS3FS also demonstrated strong performance, the proposed method achieved superior classification results while selecting fewer biomarkers. These findings indicate that adaptive graph learning and iterative sample reweighting improve the robustness and discriminative capability of leukemia biomarker discovery.

ROC Curve Analysis

To further investigate the discrimination capability of the selected biomarkers, receiver operating characteristic (ROC) curves were generated for all compared methods using both KNN and SVM classifiers as shown in Figures 2 and 3.

The ROC curves demonstrated that the proposed framework consistently achieved the largest area under the curve among the competing methods, indicating superior sensitivity-specificity tradeoff for leukemia subtype classification. The superior ROC performance indicate that the proposed adaptive structure learning framework effectively identifies biologically informative and discriminative leukemia biomarkers while maintaining robust generalization performance.

[DOI: 10.18502/ijpho.v16i3.21751]

Table I. Performance comparison of different biomarker selection methods using the KNN classifier

Method	Genes	Acc (%)	Sen (%)	Spe (%)	Pre (%)	F1 (%)	AUC-ROC
GS^{3}FS	10	94.12	90	95.83	90	90	0.9167
RS^{3}FS	7	91.18	90	91.67	81.82	85.71	0.9125
SFS	12	88.24	70	95.83	87.50	77.78	0.968
Sr-SemiDFS	8	44.12	100	20.83	34.48	51.28	0.6625
ASLCGLFS	9	94.12	100	91.67	83.33	90.91	0.9854
AGLSOFS	13	79.41	30	100	100	46.15	0.9208
SFS-SLL	15	88.24	60	100	100	75	0.7292
Proposed	8	97.06	100	95.83	90.91	95.24	1

Table II. Performance comparison of different biomarker selection methods using the SVM classifier

Method	Genes	Acc (%)	Sen (%)	Spe (%)	Pre (%)	F1 (%)	AUC-ROC
GS3FS	9	97.06	90	100	100	94.74	0.9917
RS3FS	12	97.06	100	95.83	90.91	95.24	0.9979
SFS	6	91.18	90	91.67	81.82	85.71	0.9125
Sr-SemiDFS	15	76.47	20	100	100	33.33	0.9562
ASLCGLFS	10	91.18	100	87.50	76.92	86.96	0.9854
AGLSOFS	7	94.12	90	95.83	90	90	0.9958
SFS-SLL	9	91.18	80	95.83	88.89	84.21	0.9750
Proposed	7	100	100	100	100	100	1

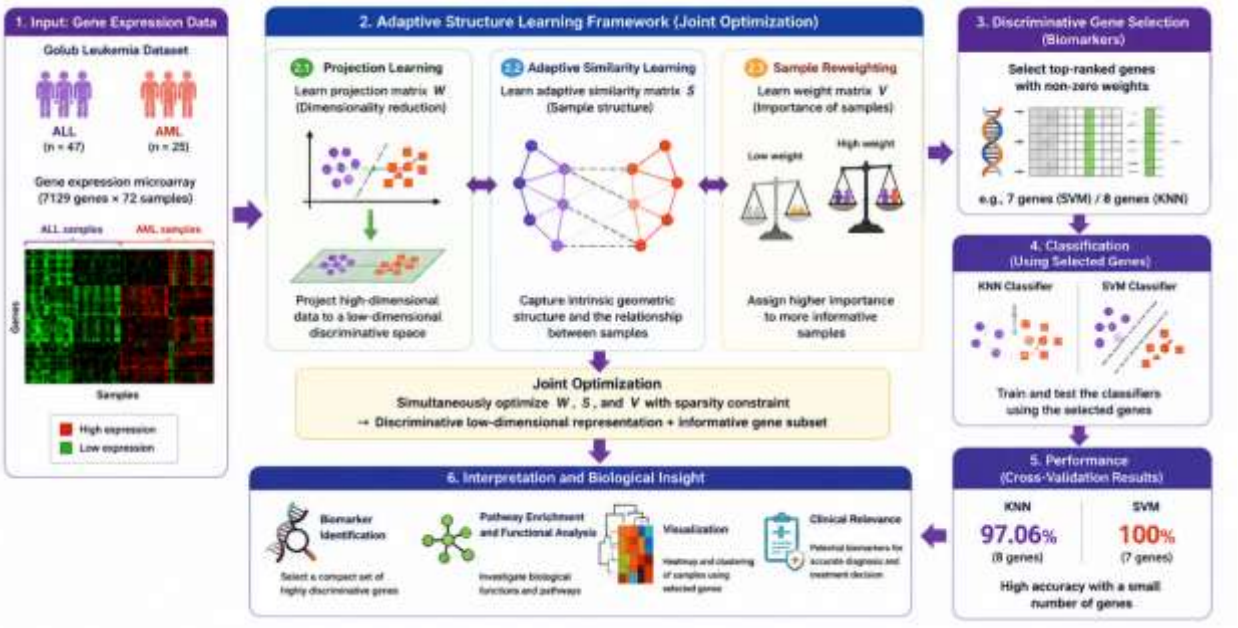

Figure 1. The overall framework of the proposed method for leukemia biomarker discovery

[DOI: 10.18502/ijpho.v16i3.21751]

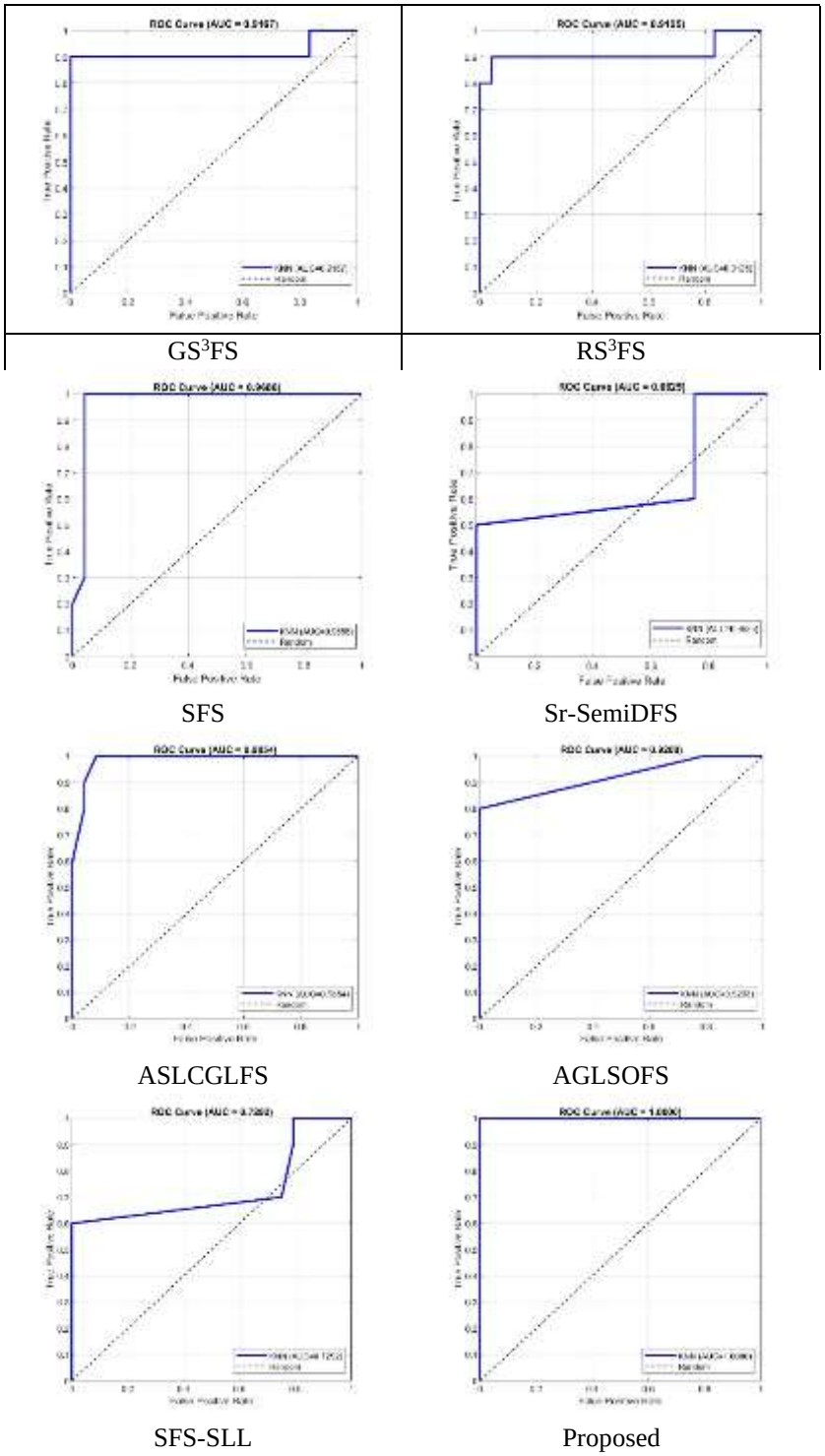

Figure 2. ROC curves of different biomarker selection methods using the KNN classifier

[DOI: 10.18502/ijpho.v16i3.21751]

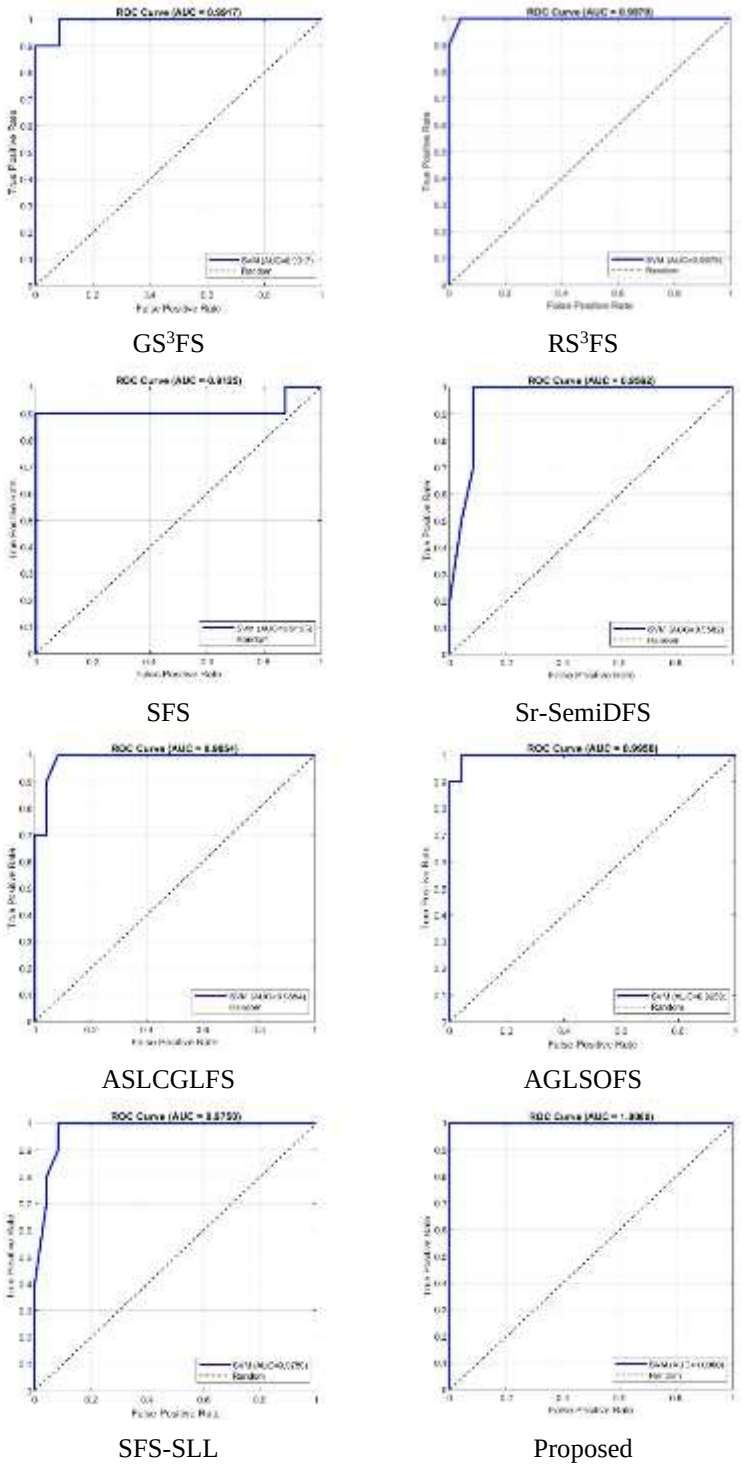

Figure 3. ROC curves of different biomarker selection methods using the SVM classifier

[DOI: 10.18502/ijpho.v16i3.21751]

Discussion

In this study, an adaptive structure learning framework was proposed for biomarker discovery from the Golub leukemia microarray dataset to distinguish ALL from AML using high-dimensional gene expression profiles. The proposed framework jointly learns the projection matrix, adaptive sample similarity structure, and sample weighting mechanism within a unified optimization model, enabling simultaneous sparse biomarker selection and adaptive structural representation of leukemia samples. By integrating adaptive graph learning with robust sparse feature selection, the proposed method aims to better capture the intrinsic geometric structure of leukemia gene expression data while reducing the influence of noisy and outlier samples.

The results demonstrate that the proposed framework achieves strong and consistent classification performance using a very small number of genes. Specifically, the proposed method achieved 97.06% classification accuracy using the KNN classifier with only 8 selected genes and obtained perfect classification performance of 100% using the SVM classifier with only 7 genes. These findings indicate that the proposed framework is capable of identifying highly discriminative biomarkers while substantially reducing the dimensionality of the original 7129-gene expression space. The ability to achieve excellent predictive performance with a compact gene subset is particularly important in clinical genomics, where smaller biomarker panels improve interpretability, reduce computational complexity, and facilitate future laboratory validation.

A growing body of literature has investigated leukemia classification using machine learning and feature selection techniques on the same or similar leukemia microarray datasets. Early studies demonstrated that appropriate gene selection substantially improves classification accuracy by removing irrelevant and redundant genes from high-dimensional expression profiles. For example, Jirapech-Umpai and Aitken utilized an evolutionary algorithm to identify the near-optimal set of predictive genes combined with classical classifiers and reported 98% classification accuracy using reduced gene subsets (13). Subsequently, fuzzy-rough set based approaches have been shown to effectively handle uncertainty in gene expression data and achieve 93.05% classification accuracy by integrating feature ranking and dependency measures (29). Ensemble learning and hybrid optimization methods have also shown promising performance for leukemia subtype prediction by combining multiple classifiers or integrating optimization algorithms with feature selection procedures (30). In addition, random forest-based guided feature selection methods have also been applied to gene expression analysis, showing stable performance across multiple high-dimensional biomedical datasets, including leukemia gene expression profiles (31). Benchmark studies further confirm that classical machine learning methods such as SVM, k-nearest neighbor, and ensemble classifiers can achieve high classification accuracy when coupled with appropriate feature selection strategies (32).

Sparse regularization-based feature selection methods have become increasingly popular for leukemia biomarker discovery due to their ability to identify compact and interpretable gene subsets. A recent study proposed a spline regression-based sparse feature selection framework (SRS3FS) using $\ell 2,p$-norm regularization with different sparsity settings for leukemia classification. The study reported a maximum accuracy of 97.06% using 10 selected genes with the SVM classifier (23). Similarly, another robust $\ell 2,p$-norm sparsity-based gene selection method achieved perfect classification performance of 100% using RF and SVM classifiers with 10 selected genes identified by the $\ell 2,1/2$-norm formulation (17). These studies demonstrated the effectiveness of sparse learning in leukemia biomarker identification; however, they relied on fixed graph structures and did not adaptively learn sample relationships during optimization.

More recent studies have explored hybrid machine learning and deep learning frameworks

[DOI: 10.18502/ijpho.v16i3.21751]

for leukemia classification. For example, a recent study utilizing leukemia gene expression data applied linear programming-based classification after selecting 25 informative genes and achieved a classification accuracy of 98.44% (30). Another study investigated six machine learning algorithms, including Gaussian Naive Bayes, Random Forest, SVM, KNN, multilayer perceptron (MLP), and logistic regression for AML and ALL classification, reporting approximately 91% accuracy for logistic regression and MLP models (33).

Advanced optimization and deep reinforcement learning methods have also recently been introduced for leukemia subtype prediction. A recent study proposed an Optimized Dueling Double Deep Q-Network (DDDQ-N) framework integrating Butterfly Optimization with Chaotic Local Search (BO-CLS) for feature selection and deep reinforcement learning for classification. The proposed framework achieved approximately 99% classification accuracy along with strong precision, recall, and F1-score values (34). Although these advanced frameworks demonstrate excellent predictive performance, they often involve multi-stage pipelines, deep architectures, and complex optimization procedures that may reduce interpretability and increase computational burden.

In contrast to many existing approaches, the proposed framework integrates adaptive similarity learning, sparse feature selection and robust sample weighting within a single optimization framework. This unified formulation provides several advantages. First, the adaptive graph learning component dynamically models local sample relationships during optimization rather than assuming a static neighborhood structure. Second, the $\ell 2,1$-norm sparsity constraint enables identification of a compact subset of highly discriminative genes. Third, the sample weighting matrix reduces the influence of noisy and potentially mislabeled samples, thereby improving robustness and stability of biomarker discovery. Together, these components allow the framework to simultaneously optimize feature relevance and structural representation of leukemia samples.

From a clinical perspective, achieving high sensitivity in leukemia subtype classification is particularly important because incorrect classification of AML patients may delay appropriate treatment and adversely affect patient outcomes. In this study, the proposed framework achieved 100% sensitivity with both KNN and SVM classifiers, indicating that all AML samples were correctly identified. This result suggests potential utility of the proposed framework as a decision-support tool for leukemia diagnosis and biomarker discovery. Furthermore, the identification of a small set of discriminative genes may facilitate future biological validation studies and development of clinically applicable molecular diagnostic panels.

Overall, the findings of this study suggest that adaptive structure learning combined with sparse feature selection provides an effective, robust, and interpretable framework for leukemia biomarker discovery. The proposed method achieves superior or competitive classification performance compared with existing machine learning and deep learning approaches while requiring fewer selected genes, highlighting its potential value for translational leukemia research and precision medicine applications.

Conclusion

The results of in this study indicated that leukemia types can be classified with very high performance using a small subset of genes selected through the proposed adaptive gene selection framework combined with machine learning algorithms. The best classification performance was achieved using the SVM classifier with only 7 selected genes, where the model obtained 100% accuracy, 100% sensitivity, 100% specificity, 100% precision, and 100% F1-score, indicating perfect classification of both AML and ALL samples without any misclassification. In addition, the model achieved an AUC-ROC value of 1.00, reflecting excellent discriminative ability between the two leukemia subtypes across all

[DOI: 10.18502/ijpho.v16i3.21751]

thresholds. These findings suggest that integrating adaptive graph learning with sparse feature selection can effectively improve the identification of clinically relevant leukemia biomarkers and may support future precision medicine applications in leukemia diagnosis.

Availability of Data

All data is available in manuscript.

Acknowledgements

None

Conflict of Interest

The authors declare that they have no conflicts of interest.

Funding

This study did not receive any specific funding.

Ethical Considerations

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

Authors' Contributions

R. Sh: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft. M. Z. S: Methodology, Validation, Writing – review & editing. A. S. H: Investigation, Validation, Writing – review & editing.

References

1. Srisuwananukorn A, Salama ME, Pearson AT. Deep learning applications in visual data for benign and malignant hematologic conditions: a systematic review and visual glossary. Haematologica 2023;108(8):1993-2010.

2. Hameed M, Raja MA, Zameer A, Dar HS, Alluhaidan AS, Aziz R. Acute myeloid leukemia classification using ReLViT and detection with YOLO enhanced by adversarial networks on bone marrow images. Sci Rep 2025;15(1):1-22.

3. Liu J, Jiang P, Lu Z, Yu Z, Qian P. Decoding leukemia at the single-cell level: clonal architecture, classification, microenvironment, and drug resistance. Exp Hematol Oncol 2024;13(1):12-39.

4. Zhang Z, Huang J, Zhang Z, Shen H, Tang X, Wu D, Bao X, Xu G, Chen S. Application of omics in the diagnosis, prognosis, and treatment of acute myeloid leukemia. Biomark Res 2024;12(1):60-103.

5. Sheikhpour R, Sarram MA, Chahooki MA, Sheikhpour R. A kernelized non-parametric classifier based on feature ranking in anisotropic Gaussian kernel. Neurocomputing 2017; 267: 545-555.

6. Barbadilla-Martínez L, Klaassen N, van Steensel B, de Ridder J. Predicting gene expression from DNA sequence using deep learning models. Nat Rev Genet 2025; 26(10): 666-680.

7. Aby AE, Salaji S, Anilkumar KK, Rajan T. A review on leukemia detection and classification using Artificial Intelligence-based techniques. Comput Electr Eng 2024; 118:109446.

8. He T, Geng J, Hou C, Li H, Zhang H, Zhao P, He P, Lu X. Predictive biomarkers and molecular subtypes in DLBCL: insights from PCD gene expression and machine learning. Discov Oncol 2025;16(1): 542-552.

9. Dubey V, Shen L. Personalized gene expression prediction in the era of deep learning: a review. Brief Bioinform 2026;27(1):1-12.

10. Cheng Y, Yang X, Wang Y, Li Q, Chen W, Dai R, Zhang C. Multiple machine-learning tools identifying prognostic biomarkers for acute myeloid leukemia. BMC Med Inform Decis Mak 2024; 24(1):1-15.

11. Alharthi R, Mehmood R, Albeshri A. A Systematic, Scalable, and Interpretable Mapping of Artificial Intelligence Research in Leukemia Using a Hybrid Machine Learning and Qualitative Framework. Electronics 2026; 15(5): 1078-1144.

12. Sánchez-Corrales YE, Pohle RV, Castellano S, Giustacchini A. Taming cell-to-cell heterogeneity in acute myeloid leukaemia with machine learning. Front Oncol 2021; 11:1-9.

[DOI: 10.18502/ijpho.v16i3.21751]

13. Jirapech-Umpai T, Aitken S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC bioinformatics 2005;6(1):148-159.

14. Ebrahimpour MK, Eftekhari M. Distributed feature selection: A hesitant fuzzy correlation concept for microarray high-dimensional datasets. Chemometr Intell Lab Syst 2018;173: 51-64.

15. Rostami M, Forouzandeh S, Berahmand K, Soltani M, Shahsavari M, Oussalah M. Gene selection for microarray data classification via multi-objective graph theoretic-based method. Artif Intell Med 2022;123: 102228.

16. Sheikhpour R, Sarram MA, Gharaghani S. Constraint score for semi-supervised feature selection in ligand-and receptor-based QSAR on serine/threonine-protein kinase PLK3 inhibitors. Chemometr Intell Lab Syst 2017;163: 31-40.

17. 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 Ped Hematol Oncol 2023; 13 (1): 13-21.

18. Alhamrani SQ, Ball GR, El-Sherif AA, Ahmed S, Mousa NO, Alghorayed SA, et al. Machine learning for multi-omics characterization of blood cancers: a systematic review. Cells 2025; 14(17):1385-1413.

19. 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.

20. Pakmanesh M, Saberi-Movahed F, Salemi A, Tiwari P. Unsupervised feature selection via graph-based proximity and structured autoencoder-like NMF. Inf Process Manag 2026; 63(6):104714.

21. Sheikhpour R, Saberi-Movahed F, Jalili M, Berahmand K. Semi-supervised feature selection with concept factorization and robust label learning. Pattern Recognit 2025:112317-112334.

22. Bishop CM, Nasrabadi NM. Pattern recognition and machine learning. New York: springer; 2006; 17:1-9.

23. Yazdanparast M, Sheikhpour R, Zangeneh Soroush M, Ghanizadeh F. In Silico Identification of Effective Genes for Acute Leukemia Classification Using a Spline Regression-based Framework. Iran J Ped Hematol Oncol 2024; 14 (2):104-115.

24. Li X, Zhang Y, Zhang R. Semisupervised feature selection via generalized uncorrelated constraint and manifold embedding. IEEE Trans Neural Netw Learn Syst 2021; 33(9):5070-5079.

25. Fan M, Zhang X, Hu J, Gu N, Tao D. Adaptive data structure regularized multiclass discriminative feature selection. IEEE Trans Neural Netw Learn Syst 2021; 33(10):5859-5872.

26. Lai J, Chen H, Li W, Li T, Wan J. Semi-supervised feature selection via adaptive structure learning and constrained graph learning. Knowl Based Syst 2022; 251: 109243-109256.

27. Liao H, Chen H, Yin T, Horng SJ, Li T. Adaptive orthogonal semi-supervised feature selection with reliable label matrix learning. Inf Process Manag 2024; 61(4):103727-103755.

28. Zhang C, Zhu L, Shi D, Zheng J, Chen H, Yu B. Semi-supervised feature selection with soft label learning. IEEE-CAA J AUTOMATIC 2022. 1-13.

29. Farahbakhshian SF, Ahvanooey MT. A new gene selection algorithm using fuzzy-rough set theory for tumor classification. arXiv preprint arXiv:2003.12386 2020. 1-10.

30. Ilyas M, Aamir KM, Manzoor S, Deriche M. Linear programming based computational technique for leukemia classification using gene expression profile. Plos one 2023;18(10): e0292172.

31. Deng H, Runger G. Gene selection with guided regularized random forest. Pattern Recognit. 2013;46(12):3483-3489.

32. Sewak MS, Reddy NP, Duan ZH. Gene expression based leukemia sub-classification using committee neural networks. Bioinform Biol Insights. 2009;3: BBI-S2908.

33. Raut R, Lokare S, Gavkare S, Narute M. Leukemia Subtype Classification via Gene Expression Using Machine Learning

[DOI: 10.18502/ijpho.v16i3.21751]

Algorithms. In 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS) 2024 Jun 28.1-7.

34. Jayakrishnan R, Meera S. Multiclass classification of leukemia cancer subtypes using gene expression data and Optimized Dueling Double Deep Q-network. Chemometr Intell Lab Syst 2025; 262:105402.

[DOI: 10.18502/ijpho.v16i3.21751]