Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2025;14(4):1054-8

Sparse gene selection and classification for acute leukemia diagnosis using a hybrid machine learning framework

Esra Pamukcu

Fırat University, Faculty of Science, Department of Statistics, Elazığ, Türkiye

Received 01 May 2025; Accepted 03 July 2025
Available online 28 October 2025 with doi: 10.5455/medscience.2025.05.114

Abstract

Acute leukemia, including Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML), presents challenges in accurate subtype classification due to the high dimensionality of gene expression data. High-dimensional gene expression datasets, such as those used in leukemia research, are frequently analyzed using machine learning methods to overcome the challenges of feature selection and classification. In recent years, hybrid approaches that combine feature selection algorithms with classification models have been increasingly employed to enhance classification performance in biomedical applications. In this study, a hybrid machine learning framework combining Least Absolute Shrinkage and Selection Operator (LASSO) regression for feature selection and Support Vector Machine (SVM) for classification was applied to the publicly available leukemia dataset, which includes gene expression profiles from 72 bone marrow samples (47 ALL, 25 AML) across 3,571 genes. LASSO regression identified the most informative genes, reducing the dataset's dimensionality, and these genes were subsequently used as input features for the SVM classifier. The classification achieved an overall accuracy of 93.33%, demonstrating the robustness of the selected features. The findings confirm the applicability of combining LASSO and SVM in gene expression-based classification tasks.

Keywords: Leukemia classification, LASSO regression, support vector machine (SVM), gene expression analysis

Introduction

Acute leukemia, encompassing Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML), stands as one of the most aggressive hematologic malignancies. Timely and precise differentiation between these subtypes is vital for effective treatment planning and enhancing patient outcomes. Traditional diagnostic methods, which rely on morphological and immunophenotypic analyses, can be subjective and time-intensive. Consequently, molecular-based classification techniques utilizing gene expression profiles have garnered significant attention in leukemia research [1,2].

Gene expression profiling offers a comprehensive snapshot of cellular activity, enabling the identification of distinct molecular signatures associated with various leukemia subtypes [3]. However, the high dimensionality of gene expression data—where the number of genes significantly surpasses the number of patient samples—poses a substantial challenge [4]. This imbalance necessitates effective feature selection methods to isolate the most informative genes, thereby improving classification performance and interpretability [5].

Machine learning approaches have emerged as powerful tools in addressing high-dimensional data challenges, particularly in biomedical applications. Among these, the Least Absolute Shrinkage and Selection Operator (LASSO) regression is a prominent technique for feature selection [6]. LASSO introduces a penalty that encourages sparsity in the model coefficients, effectively selecting a subset of relevant features while shrinking the rest toward zero [7]. This property makes LASSO especially useful in scenarios with limited sample sizes and a large number of predictors, as it enhances model generalizability and reduces overfitting. Studies have demonstrated the efficacy of LASSO in gene selection for various cancers, including leukemia [8].

CITATION
Pamukcu E. Sparse Gene Selection and Classification for Acute Leukemia Diagnosis Using a Hybrid Machine Learning Framework. Med Science. 2025;14(4):1054-8.

Corresponding Author: Esra Pamukcu, Fırat University, Faculty of Science, Department of Statistics, Elazığ, Türkiye
Email: epamukcu@firat.edu.tr

DOI: 10.5455/medscience.2025.05.114

Support Vector Machines (SVM) are another robust machine learning technique widely used for classification tasks, particularly in high-dimensional spaces. SVM aims to find the optimal hyperplane that separates classes by maximizing the margin between them. Its effectiveness in handling high-dimensional data makes it a suitable choice for classifying gene expression profiles. The combination of LASSO for feature selection and SVM for classification has been explored in various studies, demonstrating improved performance in cancer classification tasks [9].

Gene expression-based classification of leukemia subtypes has been extensively investigated in the literature. As an early contribution to the field, Golub et al. conducted a pioneering study that analyzed gene expression data to automate the classification of leukemia subtypes. Their findings demonstrated a clear correlation between gene expression patterns and leukemia types [10]. Early studies using the Golub dataset focused on applying individual machine learning methods to address the challenges posed by the high dimensionality of gene expression data. Guyon et al. (2002) further improved classification accuracy by combining SVM with Recursive Feature Elimination [11]. Zucknick et al. (2008) explored multivariate classification methods, including LASSO regression, highlighting their stability and interpretability in high-dimensional gene expression analyses [12].

In recent years, there has been a growing trend towards hybrid approaches that integrate feature selection algorithms with classification models, aiming to enhance the performance and interpretability of biomedical data analyses. Gupta and Gupta (2020) combined SVM-RFE with LASSO regression to develop a cost- and time-efficient gene expression-based diagnostic method for breast cancer [13]. Luo and Yu (2024) applied LASSO regression and SVM-RFE for identifying core diagnostic genes in sepsis, achieving high diagnostic accuracy [14]. Jiang and Jiang (2023) implemented a hybrid LASSO-SVM framework to screen gene markers for pulmonary arterial hypertension, demonstrating the efficacy of this combined approach [15]. Another noteworthy comparative analysis was done by Shadman Sakib et al. which assessed the effectiveness of various ML algorithms such as SVM, Logistic Regression (LR), Decision trees (DT), Random Forest (RF), K-Nearest Neighbours (K-NN) and Naïve Bayes after performing dimensionality reduction using Principal Component Analysis (PCA) [16]. More recently, Mahmood and Kadir (2025) demonstrated that sparsity regularization further improves gene selection and leukemia subtype classification, reinforcing the relevance of hybrid and regularized approaches in current research [17]. These recent studies demonstrate that hybrid approaches for gene expression-based classification remain a current and actively researched area.

In this study, we evaluate the performance of a hybrid machine learning approach that integrates LASSO-based gene selection with SVM classification for distinguishing between ALL and AML subtypes using microarray gene expression data. To assess the applicability of this established approach in leukemia classification, we utilized the leukemia dataset from Golub et al. (1999) [10]. This dataset comprises 72 bone marrow samples, including 47 ALL and 25 AML cases, with gene expression measured across 3.571 genes. Through this analysis, we aim to contribute to the growing body of literature demonstrating the utility of hybrid machine learning models in gene expression-based leukemia classification.

Material and Methods

LASSO Regression for Feature Selection

Least Absolute Shrinkage and Selection Operator (LASSO) regression is a widely used feature selection technique that introduces an L_1 penalty to shrink the coefficients of less relevant features to zero, effectively selecting a sparse subset of informative genes. Given a dataset with gene expression values, LASSO solves the following optimization problem:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where y_i represents the class labels (ALL or AML), x_i denotes the gene expression values, β is the regression coefficient vector, and λ is the regularization parameter controlling the sparsity of the solution [7]. In this study, λ was selected using 5-fold cross-validation based on minimizing the validation MSE that promotes a sparse solution.

Support Vector Machines (SVM)

Support Vector Machines (SVM) is a powerful classification algorithm that performs well in high-dimensional datasets. SVM aims to determine the optimal hyperplane ($w^T x+b=0$) that best separates two classes by maximizing the margin between them. The optimization problem is formulated as follows:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

$$subejct\ to\ y_i(w^T x + b) \geq 1, \quad \forall i$$

where $x_i \in \mathbb{R}^d$ represents the gene expression vector for the ith sample, y_i {-1,1} denotes the class labels, w is the weight vector, and b is the bias term [9].

In this study, we utilized a linear kernel SVM to classify Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) subtypes. The SVM model was implemented using a

DOI: 10.5455/medscience.2025.05.114

linear kernel with default parameters (C = 1.0, class_weight = None) as provided by the scikit-learn library. The genes selected via LASSO regression were input features, ensuring a reduced and informative feature space. The model's performance was evaluated using accuracy, precision, recall, and F1-score metrics, demonstrating its effectiveness in distinguishing leukemia subtypes.

Data Set and Software

The leukemia dataset consists of gene expression profiles obtained from two different tumor tissues: Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). This dataset was originally introduced by Golub et al. (1999) and has been widely used for leukemia classification and biomarker discovery [7]. The dataset includes 47 ALL patients and 25 AML patients, totaling 72 bone marrow samples. Gene expression levels were measured using Affymetrix Hgu6800 microarrays, capturing expression data for 3.571 genes across these samples. All 72 samples in the dataset were included in the analysis without any exclusion. The dataset is publicly available and fully anonymized, containing no personal or sensitive information. Therefore, no ethical approval was required. All analyses in this study were conducted using the latest version of Python (3.13.2) [18].

Results

Data Preprocessing

To ensure optimal model performance and accurate feature selection, the following preprocessing steps were applied:

1. Data Splitting: The dataset was divided into 80% training (57 samples) and 20% testing (15 samples), ensuring class balance through stratified sampling.
2. Normalization: Since gene expression data often varies in scale, standardization (zero mean, unit variance) was applied to all features before feeding them into the model.

Gene Selection via LASSO Regression

We applied Least Absolute Shrinkage and Selection Operator (LASSO) regression to identify the most relevant genes for classifying ALL and AML. The dataset was divided into an 80% training set (57 samples) and 20% test set (15 samples), and standardization was performed to ensure optimal model performance.

The LASSO regression model was trained with 5-fold cross-validation to determine the optimal λ regularization parameter. After fitting the model with λ=0.016, we identified 45 genes with nonzero coefficients, indicating their potential significance as biomarkers for leukemia classification. Figure 1 illustrates the non-zero coefficients of the 45 genes selected via LASSO regression, ranked by magnitude.

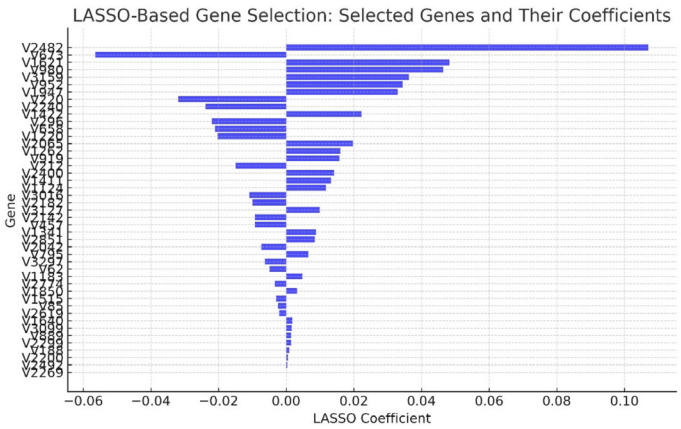

Figure 1. The bar chart visualizing the selected genes and their LASSO coefficients. The genes are sorted by the absolute value of their coefficients to highlight the most influential ones.

Classification Using SVM

After selecting the most informative genes, the classification was conducted using the Support Vector Machines (SVM) algorithm with a linear kernel function. Feature selection was performed using the LASSO method, resulting in a subset of 45 selected genes. The dataset was partitioned into 80% training and 20% testing to ensure a robust evaluation of the model's performance. The classification performance was assessed through accuracy, a confusion matrix, and the F1-score, providing a comprehensive evaluation of the model's predictive capability. The SVM classifier was implemented using a linear kernel with default hyperparameters, since it yielded high classification performance without requiring further tuning. Table 1 presents the confusion matrix showing classification performance on the test set.

Table 1. Confusion matrix of LASSO-SVM on the test set (15 samples)

True Label	Predicted ALL	Predicted AML
ALL (0)	10	0
AML (1)	1	4

(ALL): Acute lymphoblastic leukemia; (AML): Acute myeloid leukemia

The detailed classification report is:

- ALL (0): Precision = 91%, Recall = 100%, F1-score = 95%
- AML (1): Precision = 100%, Recall = 80%, F1-score = 89%
- Overall accuracy: 93.33%
- F1-score: 92%

To further investigate the discriminative power of the selected genes, a heatmap was generated to illustrate the expression levels of the 45 genes across the 15 test set samples. The heatmap provides a clear visual representation of how the gene expression patterns vary between ALL and AML patients. Genes with higher expression levels are shown in warm colors (red), while those with lower expression are depicted in cooler colors (blue). This visualization highlights the distinct expression profiles of selected genes, reinforcing their relevance in leukemia classification.

DOI: 10.5455/medscience.2025.05.114

As shown in Figure 2, the gene expression heatmap clearly differentiates between ALL and AML samples based on selected genes.

Figure 2. Heatmap visualization of the expression levels of the 45 LASSO-selected genes across the 15 test set samples.

To assess the distribution of samples in a reduced-dimensional space, Principal Component Analysis (PCA) was applied to the expression levels of the 45 selected genes across all 72 patients. The first two principal components were used to visualize the separation between ALL and AML samples. The PCA plot demonstrates that the selected genes effectively differentiate the two leukemia subtypes, as the samples cluster into distinct regions corresponding to their respective classes. This further validates the effectiveness of LASSO-based feature selection in capturing meaningful biological variation in gene expression profiles. Figure 3 displays a PCA projection of the test samples, demonstrating distinct clustering of ALL and AML cases.

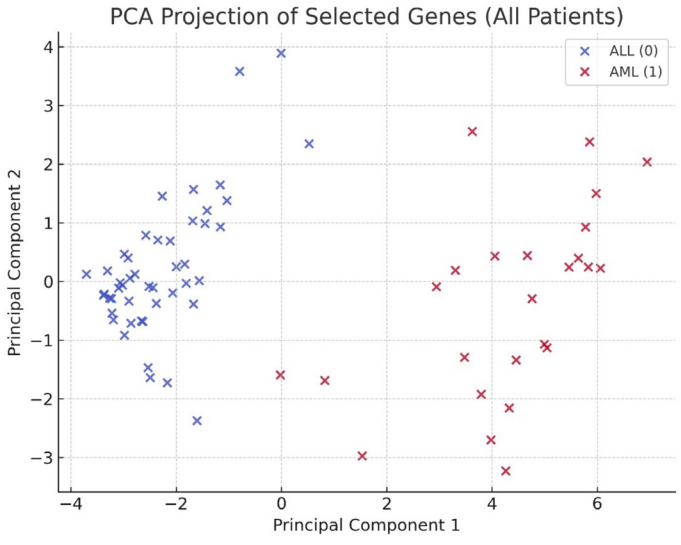

Figure 3. Principal Component Analysis (PCA) plot of the 72 patient samples based on the expression levels of the 45 LASSO-selected genes.

The results indicate that LASSO successfully identified a subset of genes that provide robust classification of leukemia subtypes, enhancing interpretability while maintaining high predictive performance. The SVM classifier, trained on these selected genes, demonstrated high accuracy, with only one misclassified AML sample.

Discussion

The findings of this study underscore the potential of integrating LASSO regression with SVM classification for leukemia subtype differentiation using gene expression data. The proposed framework addressed the high dimensionality and limited sample size challenges inherent in microarray datasets by selecting a sparse subset of informative genes. Compared to conventional gene selection techniques such as filter-based methods (e.g., t-test, ANOVA, Mutual Information), ranking approaches (e.g., Signal-to-Noise Ratio, Information Gain), and wrapper methods (e.g., Recursive Feature Elimination, Genetic Algorithms) [19], LASSO's built-in regularization mechanism provides an advantage by preventing overfitting while ensuring that only the most relevant genes contribute to classification [8]. Unlike filter and ranking methods, which do not consider gene dependencies, or wrapper methods, which can be computationally expensive and prone to overfitting in small datasets, LASSO simultaneously performs feature selection and regularization [20, 21]. LASSO provides direct interpretability via non-zero coefficients, which serve as intrinsic indicators of feature importance. This makes it particularly suitable for high-dimensional gene expression data, as it enhances model generalizability while preserving interpretability [22].

Conclusion

This study demonstrates the effectiveness of a hybrid machine learning approach that combines LASSO regression for gene selection with SVM for leukemia classification. By leveraging gene expression profiling data from Golub et al. (1999), our framework successfully identified 45 genes that distinguish between ALL and AML subtypes with 93.33% high classification accuracy. Although no formal sensitivity analysis was performed, the model produced stable results on the test set. The results highlight the advantages of LASSO-based feature selection in reducing dimensionality.

Limitations

Although transformer-based and ensemble methods (e.g., XGBoost, Random Forest) have shown high predictive power in biomedical data analysis, their applicability to small-sample, high-dimensional datasets remains limited. These models often require extensive hyperparameter tuning and large-scale data to avoid overfitting. In contrast, the LASSO-SVM framework provides both sparse feature selection and robust classification performance with transparent outputs, making it well-suited for gene expression analysis with limited samples. Additionally, while several prior studies have also applied hybrid or multivariate

models to the same dataset (e.g., SVM-RFE, PCA + classifiers [11,12,16]), direct numerical comparison was not pursued in this study due to differences in preprocessing pipelines, evaluation splits, and gene selection strategies.

The conducted methodology demonstrated a tool for improving leukemia diagnosis and facilitating biomarker discovery. However, since the gene annotations were not available in the dataset, the clinical relevance of the selected features could not be further explored. Ultimately, integrating such computational approaches with clinical diagnostics has the potential to enhance precision medicine strategies for leukemia management. The generalizability of the results is limited by the small sample size, absence of gene annotations, and lack of external validation. Future work should aim to validate these findings using independent datasets and to incorporate additional machine learning strategies, along with biological confirmation of the selected features, to enhance the generalizability and robustness of the classification framework.

Data Availability Statement

The data that support the findings of this study are publicly available and originate from the Golub et al. (1999) leukemia dataset with reference number 10.

Code and Methods

The analyses were performed using Python. No custom code was developed or shared since the study relied on standard statistical procedures implemented in these software environments.

Conflict of Interests

The authors declare that there is no conflict of interest in the study.

Financial Disclosure

The authors declare that they have received no financial support for the study.

Ethical Approval

The study used openly accessible data, and no human or sensitive data were involved. Therefore, ethical approval was not required.

References

1. Duncavage EJ, Bagg A, Hasserjian RP, et al. Genomic profiling for clinical decision making in myeloid neoplasms and acute leukemia. Blood. 2022;140:2228-47.
2. Eckardt JN, Stölzel F, Kunadt D, et al. Molecular profiling and clinical implications of patients with acute myeloid leukemia and extramedullary manifestations. J Hematol Oncol. 2022;15:60.
3. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med. 2004;350:1617-28.
4. Clark AJ, Lillard JW Jr. A comprehensive review of bioinformatics tools for genomic biomarker discovery driving precision oncology. Genes. 2024;15:1036.
5. Baali S, Hamim M, Moutachaouik H, Hain M, EL Moudden I. Feature Selection for High-Dimensional Gene Expression Data: A Review. In: International Conference on Smart Applications and Data Analysis. Cham: Springer Nature Switzerland; 2024:74-92.
6. Xiong Y, Ling QH, Han F, Liu QH. An efficient gene selection method for microarray data based on LASSO and BPSO. BMC Bioinformatics. 2019;20:1-13.
7. Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc Series B Stat Methodol. 1996;58:267-88.
8. Ma S, Song X, Huang J. Supervised group Lasso with applications to microarray data analysis. BMC Bioinformatics. 2007;8:1-17.
9. Vapnik V. Support-vector networks. Mach Learn. 1995;20:273-97.
10. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286:531-7.
11. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389-422.
12. Zucknick M, Richardson S, Stronach EA. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. Int J Biostat. 2008;4:Article 27.
13. Gupta M, Gupta B. A novel gene expression test method of minimizing breast cancer risk in reduced cost and time by improving SVM-RFE gene selection method combined with LASSO. J Integr Bioinform. 2020;17:20200027.
14. Luo L, Yu J. Identifying core genes in sepsis by LASSO regression and SVM-RFE algorithm. Arch Med Sci. 2024;20:511-20.
15. Jiang C, Jiang W. Lasso algorithm and support vector machine strategy to screen pulmonary arterial hypertension gene diagnostic markers. Scott Med J. 2023;68:21-31.
16. Sakib S, Tanzeem AK, Tasawar IK, et al. Blood cancer recognition based on discriminant gene expressions: A comparative analysis of optimized machine learning algorithms. Presented at an International Conference; 2021.
17. Mahmood NH, Kadir DH. Sparsity regularization enhances gene selection and leukemia subtype classification via logistic regression. Leuk Res. 2025;107663.
18. Python Software Foundation. Python 3.13.2 Documentation. Available at: https://www.python.org/downloads/. Accessed 2025.
19. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157-82.
20. Almutiri T, Saeed F. Review on feature selection methods for gene expression data classification. In: Emerging Trends in Intelligent Computing and Informatics: Data Science, Intelligent Information Systems and Smart Computing 4. Springer International Publishing; 2020:24-34.
21. Beder T, Hansen BT, Hartmann AM, et al. The gene expression classifier ALLCatchR identifies B-cell precursor ALL subtypes and underlying developmental trajectories across age. Hemasphere. 2023;7:e939.
22. Covert I, Gala R, Wang T, et al. Predictive and robust gene selection for spatial transcriptomics. Nat Commun. 2023;14:2091.