Proceedings of ICBioMed 2026 Symposium: Biomedical Engineering and Smart Medicine
DOI: 10.54254/2753-8818/2026.HZ37182

Identification of Candidate Diagnostic Biomarkers and Construction of an Interpretable Machine-Learning Model for Acute Myeloid Leukemia

Shan Liu

Haide College, Ocean University of China, Qingdao, China
23220001034@stu.ouc.edu.cn

Abstract. Acute myeloid leukemia (AML) is a molecularly heterogeneous disease, with considerable variation in molecular abnormalities and gene-expression patterns among patients. This study aimed to identify a compact diagnostic gene signature and develop an interpretable machine-learning model for distinguishing AML from normal samples. Gene-expression data from GSE9476, comprising 26 AML and 38 normal samples, were divided into a 51-sample training set and an independent 13-sample test set. Training-restricted differential analysis identified 283 significant genes. Least absolute shrinkage and selection operator regression, combined with Random Forest screening, reduced this set to ALDH1A1, GPX1, CAPRIN2, MFSD10 and IL3RA. Logistic Regression, linear support vector machine and Random Forest models were compared by five-fold stratified cross-validation. Logistic Regression was retained as the primary model and achieved a test ROC-AUC of 1.000, with 92.3% accuracy, 80.0% sensitivity and 100% specificity. Shapley additive explanations ranked GPX1, ALDH1A1 and IL3RA as the strongest contributors, with effects aligned with the differential expression results. External validation in 77 GSE30029 samples produced a ROC-AUC of 0.969 (95% confidence interval: 0.922–1.000), 92.2% accuracy and 100% specificity. Agreement across independent cohorts and microarray platforms supports the five-gene panel as a reproducible candidate signature. Larger multicenter cohorts and experimental assays remain necessary before clinical application.

Keywords: acute myeloid leukemia, diagnostic biomarkers, gene expression, interpretable machine learning, SHAP

1. Introduction

Acute myeloid leukemia (AML) is an aggressive hematological malignancy in which immature myeloid cells expand and disturb normal blood formation. Yet the disease does not follow one clear clinical course. Patients with similar cell morphology may carry different molecular changes, respond differently to treatment, and have very different outcomes. Diagnosis therefore combines morphology, immunophenotyping, cytogenetics, and molecular findings [1]. Even with this broader view, finding clear disease-related molecular patterns remains difficult.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.HZ37182

This difficulty reflects the biological diversity of AML. Genomic studies have reported recurrent mutations, chromosomal changes, and epigenetic alterations, but no single abnormality describes the whole disease [2]. Gene-expression data provide a different view by showing the activity of many genes at the same time. Earlier studies found that expression patterns could separate AML samples into groups with biological and clinical meaning [3]. Public microarray repositories now allow these patterns to be studied in existing cohorts and compared across independent datasets.

Such data are valuable, but they are not easy to turn into a practical diagnostic signature. A typical microarray measures thousands of genes in a much smaller number of samples. A model may therefore learn random noise or features that are specific to one cohort. Its performance can fall when the patient group or laboratory platform changes. Feature selection helps reduce this problem, although a high accuracy score alone is not enough. For biomedical use, researchers also need to know which genes guide a prediction and whether the same pattern appears in another dataset. A small panel with visible gene effects is easier to examine than a large model that gives little explanation for its output.

The present study used GSE9476 as the development cohort and divided its 64 samples into a 51-sample training set and an untouched 13-sample test set. Differential expression analysis and feature selection were limited to the training data to reduce information leakage. LASSO regression and random forest importance were combined to identify a compact gene panel. Logistic regression, linear support vector machine, and random forest classifiers were then compared by stratified cross-validation. SHAP analysis was used to show how the selected genes affected individual predictions. GSE30029 provided external validation on a different microarray platform. Rather than focusing only on internal accuracy, the study aimed to find a small AML-related signature that was understandable and could retain useful discrimination in an independent cohort.

2. Materials and methods

2.1. Data acquisition and preprocessing

Gene-expression profiles were obtained from the Gene Expression Omnibus (GEO) database [4]. GSE9476, generated on the Affymetrix Human Genome U133A Array (GPL96), served as the development cohort [5] and contained 26 acute myeloid leukemia (AML) samples and 38 normal hematopoietic samples. Normalized expression values and GPL96 annotations were parsed from the downloaded files. Probes without a valid gene symbol were discarded, while probes mapped to the same symbol were averaged, producing a matrix of 13,299 genes across 64 samples. Stratified sampling with a random seed of 42 assigned 21 AML and 30 normal samples to the training set; the remaining 5 AML and 8 normal samples formed an independent test set. Expression distributions were examined across samples, and principal component analysis was performed on the 1,000 genes with the highest variance after mean centering. GSE30029, measured on the Illumina HumanHT-12 V3.0 platform (GPL6947), was reserved for external validation [6].

2.2. Differential expression and feature selection

Differential expression analysis was confined to the 51 training samples so that the internal test set did not influence feature screening. Linear models and empirical Bayes moderation were implemented with limma [7]. The contrast was defined as AML minus normal, and P values were adjusted by the Benjamini–Hochberg procedure. Genes with an adjusted P value below 0.05 and an absolute log2 fold change above 1 were considered differentially expressed. Ambiguous multi-gene

DOI: 10.54254/2753-8818/2026.HZ37182

symbols were removed before machine-learning analysis. Feature selection combined least absolute shrinkage and selection operator (LASSO) regression with random forest importance [8, 9]. LASSO used standardized predictors, class-balanced logistic regression and five-fold stratified cross-validation. Fifty logarithmically spaced C values from 10^{-4} to 10^{4} were evaluated by receiver operating characteristic area under the curve (ROC-AUC). Random forest importance was estimated from 1,000 trees with square-root feature sampling and a minimum leaf size of two. The final panel consisted of genes retained by LASSO and ranked among the 20 most important random forest features.

2.3. Model development and interpretation

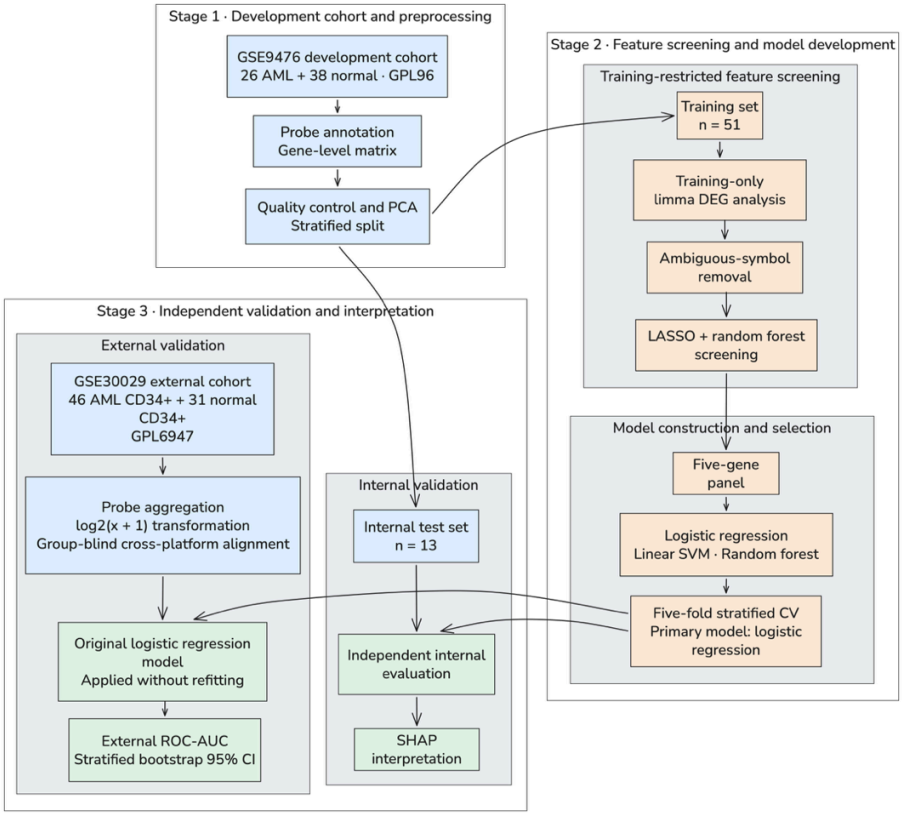

Figure 1. Analytical workflow for biomarker screening, model development, interpretation, and external validation

Logistic regression, linear support vector machine (SVM) and random forest classifiers were implemented in Python with scikit-learn [10]. Standardization was fitted within the logistic regression and SVM pipelines, and class-balanced learning was used for all three models. Model comparison relied on five-fold stratified cross-validation of the training set. Mean cross-validation ROC-AUC guided selection of the primary classifier; a tied top score was resolved in favor of the compact probabilistic logistic regression model. Internal test performance was not used for model selection. Each classifier was evaluated on the untouched 13-sample test set using ROC-AUC, accuracy, sensitivity, specificity, precision, F1 score and the confusion matrix. Predictions used a probability threshold of 0.5. The selected model was interpreted with Shapley additive explanations (SHAP) [11]. A permutation-based explainer used the 51 training samples as background data and

DOI: 10.54254/2753-8818/2026.HZ37182

calculated SHAP values for all 13 test samples. Global importance was represented by the mean absolute SHAP value for each gene.

2.4. External validation and statistical analysis

External validation used 46 AML CD34+ and 31 normal CD34+ samples selected from GSE30029 [6]. Candidate genes were matched to GPL6947 probe annotations, and multiple probes assigned to one gene were averaged. Background-corrected values below zero were truncated to zero before log2(x + 1) transformation. Without using class labels, each external gene was converted to z scores and mapped to the mean and standard deviation stored in the development model. The original logistic regression pipeline was applied without refitting. Discrimination was measured by ROC-AUC; accuracy, sensitivity, specificity and the confusion matrix were calculated at the prespecified threshold of 0.5. The 95% confidence interval for external ROC-AUC was estimated through 10,000 iterations of stratified percentile bootstrap sampling. AML and normal cases were resampled separately with replacement, preserving the original class sizes. A random seed of 42 was maintained throughout data splitting, feature selection, model fitting and bootstrap analysis. The complete analytical workflow is summarized in Figure 1.

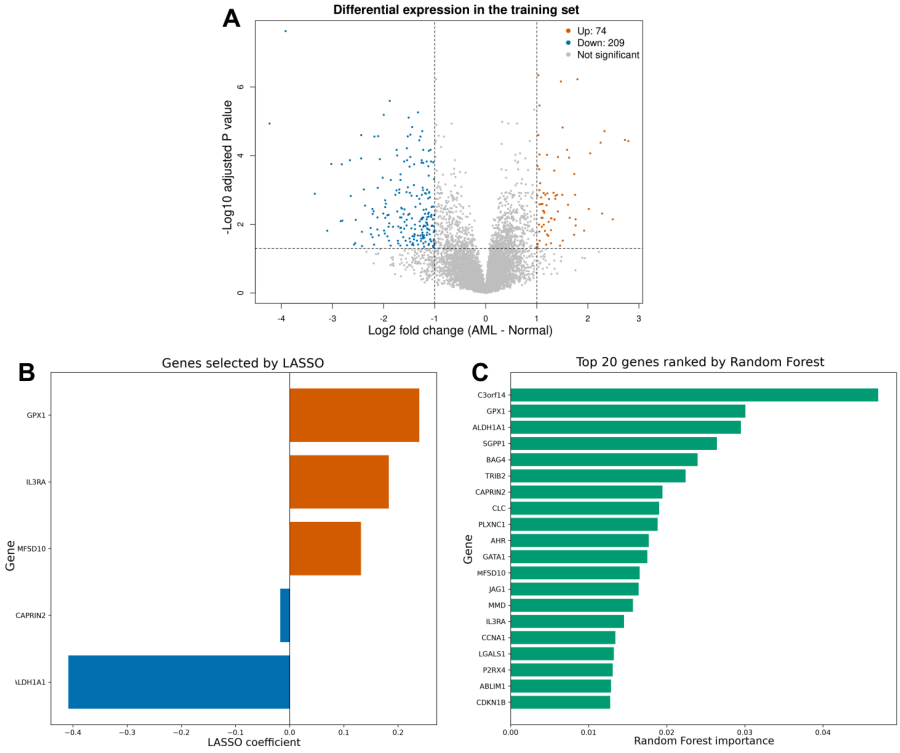

Figure 2. Differential expression and feature selection in the GSE9476 training cohort. (A) Volcano plot of genes identified by limma using an adjusted P value below 0.05 and an absolute log2 fold change above 1. (B) Nonzero coefficients of genes retained by LASSO regression. (C) The 20 genes with the highest random forest importance

DOI: 10.54254/2753-8818/2026.HZ37182

3. Results

3.1. Dataset characteristics and differential expression

The processed GSE9476 matrix contained 13,299 genes from 64 samples, including 26 AML and 38 normal controls. Median expression values ranged narrowly from 4.384 to 4.631, while sample means varied between 4.969 and 5.139. No missing or non-finite values were detected. Principal component analysis revealed pronounced sample-type structure: AML samples occupied a broad central region, while normal bone marrow, peripheral blood and peripheral blood stem-cell samples formed several tighter clusters. The first two components explained 46.7% and 12.0% of the total variance. Limma analysis of the 51-sample training set identified 283 differentially expressed genes, comprising 74 upregulated and 209 downregulated genes in AML. ALDH1A1 showed a pronounced negative change among the leading candidates (log2 fold change = −3.918), while GPX1 was strongly significant and upregulated (log2 fold change = 1.030). The asymmetric distribution of upregulated and downregulated genes is visible in Figure 2a.

3.2. Candidate genes and internal model performance

Removal of 11 ambiguous multi-gene annotations reduced the feature pool from 283 to 272 genes. LASSO reached its highest cross-validation ROC-AUC at C = 0.08685 and retained five non-zero coefficients. All five genes also appeared among the 20 most important Random Forest features, leaving ALDH1A1, GPX1, CAPRIN2, MFSD10 and IL3RA as the final panel. Their expression directions were internally consistent: GPX1, MFSD10 and IL3RA were upregulated in AML, whereas ALDH1A1 and CAPRIN2 were downregulated. The differential-expression and feature-selection results are presented in Figure 2.

As shown in Table 1, Logistic Regression and linear SVM both achieved a mean five-fold cross-validation ROC-AUC of 1.000 with no variation across folds. Random Forest produced a mean value of 0.993 ± 0.013. All three classifiers reached a test ROC-AUC of 1.000. Logistic Regression and SVM each classified 12 of 13 test samples correctly, corresponding to 92.3% accuracy, 80.0% sensitivity and 100% specificity. Random Forest correctly classified all 13 samples. Because model selection was based on training cross-validation rather than test performance, Logistic Regression was retained as the primary classifier after attaining the joint-highest cross-validation score.

Table 1. Cross-validation and internal test performance of the evaluated classifiers

Model	CV ROC-AUC	Test ROC-AUC	Accuracy	Sensitivity	Specificity	Precision	F1 score
Logistic Regression	1.000 ± 0.000	1.000	0.923	0.800	1.000	1.000	0.889
Linear SVM	1.000 ± 0.000	1.000	0.923	0.800	1.000	1.000	0.889
Random Forest	0.993 ± 0.013	1.000	1.000	1.000	1.000	1.000	1.000

3.3. SHAP interpretation and external validation

SHAP analysis showed that each member of the five-gene panel contributed meaningfully to the Logistic Regression output. GPX1 had the highest mean absolute SHAP value (0.1009), closely followed by ALDH1A1 (0.0979) and IL3RA (0.0923). CAPRIN2 and MFSD10 contributed mean absolute values of 0.0769 and 0.0707. Higher GPX1, IL3RA and MFSD10 expression generally shifted predictions toward AML. Lower ALDH1A1 and CAPRIN2 expression produced a similar

DOI: 10.54254/2753-8818/2026.HZ37182

effect, matching the directions observed in the differential expression analysis. The global ranking and sample-level distribution of these effects are presented in Figure 3.

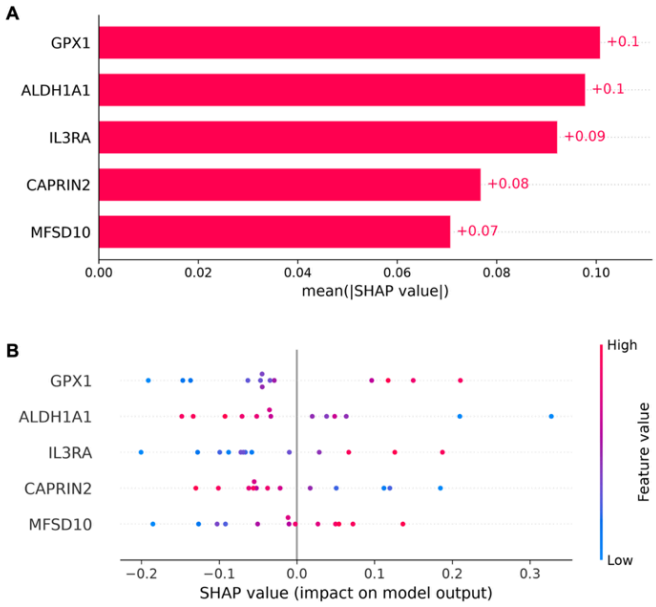

Figure 3. SHAP-based interpretation of the five-gene logistic regression model. (A) Global feature importance ranked by mean absolute SHAP value. (B) Distribution of gene-level SHAP values across the internal test samples, with color indicating the relative feature value

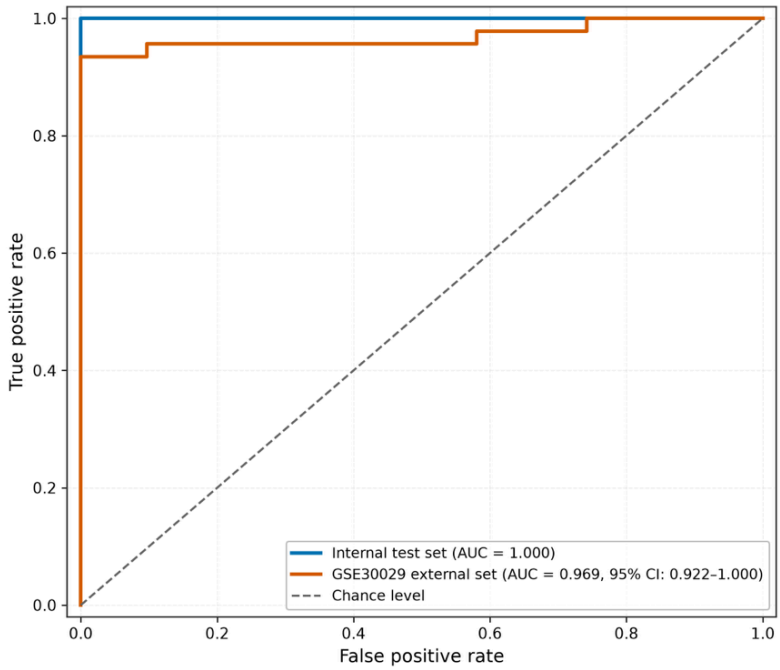

Figure 4. Receiver operating characteristic curves for internal and external validation of the five-gene logistic regression model

The original Logistic Regression model maintained strong discrimination in the independent GSE30029 cohort. Among 77 external samples, the model achieved a ROC-AUC of 0.969, with a

DOI: 10.54254/2753-8818/2026.HZ37182

bootstrap 95% confidence interval of 0.922–1.000. At the 0.5 threshold, 71 samples were correctly classified, giving an accuracy of 92.2%, sensitivity of 87.0% and specificity of 100%. All 31 normal samples were classified correctly, while 40 of 46 AML samples were identified and six were missed. The AML-related expression direction of every candidate gene agreed with the development cohort despite the different microarray platform. Figure 4 contrasts the perfect internal test ROC-AUC with the slightly lower, but stable, external performance.

4. Discussion

4.1. Biological interpretation of the five-gene panel

The five-gene panel captured a coherent expression pattern rather than a collection of unrelated statistical signals. GPX1, IL3RA and MFSD10 were elevated in AML, whereas ALDH1A1 and CAPRIN2 were reduced. The same directions appeared in GSE30029, despite differences in platform and sample composition. This agreement lends biological credibility to the model and reduces the likelihood that the panel reflects a single-cohort artifact.

Several genes also fit established features of AML biology. IL3RA encodes CD123, a surface marker enriched on AML blasts and leukemia stem-cell populations [12]. Its positive contribution to AML probability in Figure 3 is therefore biologically plausible. ALDH1A1 showed the strongest negative coefficient in the panel. Lower ALDH1A1 RNA expression has been associated with a distinct favorable-risk AML group, suggesting that its variation may carry information about AML state and heterogeneity rather than serving as a simple lineage marker [13]. GPX1 produced the highest mean absolute SHAP value and was upregulated in both cohorts. Previous work has linked GPX1-related expression patterns to poor survival, oxidative stress and an immunosuppressive microenvironment in AML [14]. Abnormal glutathione metabolism has also been described as a metabolic vulnerability of primary AML cells, with GPX1 participating in this redox network [15]. CAPRIN2 and MFSD10 have received less attention in AML. Their reproducible expression directions make them useful data-driven candidates, although their mechanistic roles require direct investigation.

4.2. Model performance and potential clinical relevance

An ROC-AUC of 1.000 in the 13-sample internal test set is encouraging, but the external cohort offers a more realistic view. In GSE30029, the logistic regression model achieved a ROC-AUC of 0.969 (95% CI: 0.922–1.000). It correctly classified all 31 normal samples but missed six of 46 AML cases. The model therefore separated the two groups well, though it is not ready to serve as a stand-alone diagnostic tool.

The five-gene panel is also easier to understand than a large signature. SHAP showed how each gene shifted predictions toward AML or normal. This fits earlier evidence that gene-expression data can help classify AML despite its biological diversity [1-3]. The panel may later support screening, but it should complement morphology, flow cytometry, cytogenetics and molecular testing.

4.3. Limitations and future directions

Sample size is the main limitation. GSE9476 contained 64 samples, with only 13 used for internal testing. Training-only feature selection reduced data leakage, yet selecting five genes from

DOI: 10.54254/2753-8818/2026.HZ37182

thousands may still overstate performance. Normal samples also came from different hematopoietic sources, so some expression changes may reflect cell composition.

GSE30029 provided independent validation, but it used CD34+-enriched cells and a different platform. Alignment reduced platform differences without removing them fully. The study did not examine clinical variables, AML subtypes or treatment outcomes, and no experimental validation was carried out. The panel should now be tested in larger multicenter cohorts and by qPCR or protein-level assays. CAPRIN2 and MFSD10 also need further biological study. Until these steps are completed, the five genes should be treated as a candidate signature rather than a clinical diagnostic test.

5. Conclusion

A compact five-gene panel comprising ALDH1A1, GPX1, CAPRIN2, MFSD10 and IL3RA was identified through training-restricted differential analysis and complementary LASSO and Random Forest selection. The Logistic Regression model achieved a ROC-AUC of 1.000 in the internal test set and 0.969 in GSE30029, while SHAP analysis revealed how each gene influenced AML predictions. Agreement across two microarray platforms supports the reproducibility of the signature, although the small retrospective cohorts and cross-platform processing limit immediate clinical application. Validation in larger multicenter populations, followed by quantitative PCR or protein-level assessment, is required before clinical use. The transparent modelling framework provides a practical basis for further AML biomarker development.

References

[1] Döhner, H., Wei, A. H., Appelbaum, F. R., et al. (2022). Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood, 140*(12), 1345–1377. https://doi.org/10.1182/blood.2022016867.

[2] Cancer Genome Atlas Research Network. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England Journal of Medicine, 368*(22), 2059–2074. https://doi.org/10.1056/NEJMoa1301689

[3] Bullinger, L., Döhner, K., Bair, E., Fröhling, S., Schlenk, R. F., Tibshirani, R., Döhner, H., & Pollack, J. R. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *The New England Journal of Medicine, 350*(16), 1605–1616. https://doi.org/10.1056/NEJMoa031046

[4] Barrett, T., Wilhite, S. E., Ledoux, P., et al. (2013). NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Research, 41*(Database issue), D991–D995. https://doi.org/10.1093/nar/gks1193

[5] Stirewalt, D. L., Meshinchi, S., Kopecky, K. J., et al. (2008). Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes, Chromosomes and Cancer, 47*(1), 8–20. https://doi.org/10.1002/gcc.20500

[6] de Jonge, H. J. M., Woolthuis, C. M., Vos, A. Z., et al. (2011). Gene expression profiling in the leukemic stem cell-enriched CD34+ fraction identifies target genes that predict prognosis in normal karyotype AML. *Leukemia, 25*(12), 1825–1833. https://doi.org/10.1038/leu.2011.172

[7] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), e47. https://doi.org/10.1093/nar/gkv007

[8] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[9] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[11] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

DOI: 10.54254/2753-8818/2026.HZ37182

[12] Jordan, C. T., Upchurch, D., Szilvassy, S. J., et al. (2000). The interleukin-3 receptor alpha chain is a unique marker for human acute myelogenous leukemia stem cells. *Leukemia, 14*(10), 1777–1784. https://doi.org/10.1038/sj.leu.2401903

[13] Dancik, G. M., Voutsas, I. F., & Vlahopoulos, S. (2022). Lower RNA expression of ALDH1A1 distinguishes the favorable risk group in acute myeloid leukemia. *Molecular Biology Reports, 49*(4), 3321–3331. https://doi.org/10.1007/s11033-021-07073-7

[14] Zhang, J., Peng, Y., He, Y., et al. (2022). GPX1-associated prognostic signature predicts poor survival in patients with acute myeloid leukemia and involves in immunosuppression. *Biochimica et Biophysica Acta—Molecular Basis of Disease, 1868*(1), 166268. https://doi.org/10.1016/j.bbadis.2021.166268

[15] Pei, S., Minhajuddin, M., Callahan, K. P., et al. (2013). Targeting aberrant glutathione metabolism to eradicate human acute myelogenous leukemia cells. *Journal of Biological Chemistry, 288*(47), 33542–33558. https://doi.org/10.1074/jbc.M113.511170