Research Article *Open Access*

Journal of Biomedical Advancement Scientific Research

ISSN: 3069-0854

DOI: doi.org/10.63721/26JBASR0160

An AI-Driven Pipeline for High-Dimensional Gene Expression Analysis in Leukemia Classification (ALL vs AML)

Georgios Vasileiadis

Department of Tourism Economics and Management, University of the Aegean, Greece

Citation: *Georgios Vasileiadis (2026) An AI-Driven Pipeline for High-Dimensional Gene Expression Analysis in Leukemia Classification (ALL vs AML). J. of Bio Adv Sci Research, 2(5):01-14. WMJ/JBASR-160*

Abstract

High-dimensional gene expression datasets present major analytical challenges in biomedical research because the number of variables greatly exceeds the number of available samples. This study proposes an artificial intelligence–driven analytical pipeline for the classification of Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) using microarray gene expression data from the Golub leukemia dataset. The aim of the study is to develop a robust, interpretable and leakage-free machine learning framework capable of supporting precision medicine and clinical decision-making. The dataset included 7,129 genes and 72 patient samples (47 ALL and 25 AML cases). An 80/20 train–test split was applied while preserving class proportions. Initially, variance filtering retained the 5,000 most informative genes from the training set. Differential expression analysis was then performed using the limma framework with Benjamini–Hochberg False Discovery Rate correction, identifying 734 statistically significant genes (FDR ≤ 0.05). Subsequently, supervised Principal Component Analysis was conducted on the selected genes, with the first principal component explaining approximately 41% of the total variance. The resulting components were used as inputs for ridge logistic regression with internal cross-validation. The proposed pipeline achieved excellent classification performance, with test Accuracy = 1.00, Sensitivity = 1.00, Specificity = 1.00, and AUC = 1.00, while 5-fold cross-validation produced a mean AUC of 0.983. Furthermore, permutation testing generated a mean AUC close to 0.50, confirming that the observed performance was not due to random chance or data leakage. Overall, the findings demonstrate that integrating statistical feature selection, supervised dimensionality reduction and regularized machine learning can provide highly accurate and interpretable models for leukemia classification, highlighting the growing role of artificial intelligence in precision oncology and public health informatics.

*__Corresponding author__: Georgios Vasileiadis, Department of Tourism Economics and Management, University of the Aegean, Greece.

Submitted: 19.08.2026 **Accepted**: 25.08.2026 **Published**: 05.09.2026

Keywords: Machine Learning, Gene Expression Analysis, Leukemia Classification, Bioinformatics, Dimensionality Reduction, Precision Medicine

Introduction

Leukemia is a heterogeneous group of hematological malignancies characterized by the uncontrolled proliferation of abnormal white blood cells in the bone marrow and peripheral blood. Among the major leukemia subtypes, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) represent biologically and clinically distinct diseases that require different therapeutic strategies and prognostic approaches. Accurate discrimination between these two forms of leukemia is therefore essential for appropriate treatment selection, patient stratification, and precision medicine–oriented clinical decision-making (Short et al., 2024). Traditional diagnostic approaches rely on morphology, immunophenotyping, cytogenetics, and molecular testing; however, advances in transcriptomics and high-throughput genomic technologies have demonstrated that gene expression profiling can provide highly informative molecular signatures capable of distinguishing leukemia subtypes with remarkable accuracy (Sartori et al., 2025).

The rapid development of microarray and next-generation sequencing technologies has generated extremely high-dimensional biomedical datasets in which the number of variables (genes) greatly exceeds the number of available patient samples (p≫np \ gg np≫n). This characteristic creates substantial methodological and statistical challenges, including multicollinearity, overfitting, unstable parameter estimation, and increased risk of information leakage during model evaluation. Consequently, the development of robust and interpretable machine learning pipelines specifically adapted to omics data has become a major research priority in bioinformatics, computational oncology, and public health informatics (Lin et al., 2015).

Gene expression analysis has played a central role in the evolution of precision oncology because it enables the identification of molecular patterns associated with disease subtype, prognosis, therapeutic response, and biological pathways involved in carcinogenesis. In leukemia research specifically, transcriptomic profiling has revealed profound differences in the expression patterns of lymphoid and myeloid malignancies, supporting both disease classification and biomarker discovery (Cruz-Miranda et al., 2024). However, because thousands of genes are tested simultaneously, rigorous statistical control of false discoveries is required in order to avoid spurious findings and ensure reproducibility. For this reason, differential expression frameworks such as limma and False Discovery Rate (FDR) correction methods remain standard tools in high-dimensional transcriptomic studies (Storey, 2025).

At the same time, dimensionality reduction techniques are essential for transforming extremely complex genomic datasets into compact and interpretable representations. Principal Component Analysis (PCA) is among the most widely used approaches for this purpose because it summarizes the dominant variability structure of the data into a small number of orthogonal dimensions. When PCA is applied after supervised gene selection, the resulting principal components inherit biologically relevant discriminatory information, improving both visualization and downstream predictive modeling. Such supervised dimensionality reduction approaches are particularly valuable in cancer genomics, where strong molecular signals are often embedded within a very high-dimensional feature space (Wang et al., 2025).

In addition to dimensionality reduction, regularized machine learning methods are critical in small-sample genomic applications. Ridge logistic regression is especially suitable for transcriptomic classification problems because it stabilizes parameter estimation in the presence of correlated predictors and reduces the risk of overfitting. Furthermore, internal cross-validation procedures and leakage-free evaluation strategies are necessary to obtain realistic estimates of predictive performance and avoid overly optimistic results, which remain a common issue in biomedical artificial intelligence research (Apicella et al., 2025).

The aim of the present study is to present a complete high-dimensional gene expression analysis pipeline for the classification of Acute Lymphoblastic Leukemia

(ALL) and Acute Myeloid Leukemia (AML), utilizing statistical and machine learning techniques specifically designed for genomic datasets. The proposed framework integrates:

- statistical differential expression analysis on the training set,
- feature selection based on multiple testing procedures and False Discovery Rate (FDR) control,
- supervised PCA applied to selected genes so that the principal components inherit supervised information,
- classification using ridge logistic regression on the principal components, and
- evaluation through confusion matrix analysis, ROC/AUC metrics, k-fold cross-validation without information leakage, and permutation-based sanity checks (Tsega & Mullualem, 2026).

The study seeks to demonstrate how the integration of statistical genomics, dimensionality reduction, and interpretable artificial intelligence methods can produce robust and biologically meaningful predictive models for leukemia classification, while simultaneously addressing methodological concerns related to reproducibility, overfitting and data leakage in high-dimensional biomedical research.

Materials and Methods

Data and Experimental Design

The Golub microarray gene expression dataset was used in this study, consisting of:

- 7,129 genes and 72 samples.
- Two classes: ALL (n = 47) and AML (n = 25).

The dataset was divided using an 80/20 train–test split while preserving class proportions:

- Training set: 58 samples (ALL = 38, AML = 20).
- Test set: 14 samples (ALL = 9, AML = 5).

Stratified sampling reduces the risk of evaluation bias, particularly in datasets with imbalanced class distributions, and improves the reliability of predictive performance estimation in biomedical classification tasks (Araújo et al., 2024). Furthermore, maintaining a strict separation between training and testing data is considered essential in high-dimensional bioinformatics pipelines to avoid optimistic bias and data leakage (Jha & Hasija, 2026).

Preprocessing

In gene expression datasets, many genes exhibit very low variance across samples and therefore contribute mainly noise rather than discriminative biological information. For this reason, a variance filtering procedure was applied, retaining only the 5,000 genes with the highest variance exclusively on the training set, followed by alignment of the test set to the same selected genes (Ali et al., 2025).

For each gene g, the sample variance in the training set is defined as:

$$s_g^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{g,i} - \bar{x}_g\right)^2, \bar{x}_g = \frac{1}{n}\sum_{i=1}^{n} x_{g,i}$$

The retained subset of genes is therefore defined as:

$$\mathcal{G}_{top} = \arg\operatorname{top}_g(s_g^2), \mid \mathcal{G}_{top} \mid = 5000$$

It should be emphasized that the variance filter was applied only to the training data. Consequently, the test distribution was never inspected during feature selection, thereby reducing the risk of information leakage and preserving the validity of the downstream evaluation process (Apicella et al., 2025). Variance-based filtering is widely used in transcriptomic machine learning workflows because it improves computational efficiency and removes biologically uninformative features prior to downstream supervised modeling (Wang et al., 2025).

Exploratory Data Analysis

For an initial visual understanding of the data structure, a heatmap was created using the top-variance genes from the training set. In this figure, we observe the hierarchical heatmap generated from the 50 genes with the highest variance in the training dataset. Heatmaps are widely used in transcriptomic analysis because they allow simultaneous visualization of both gene expression patterns and relationships among samples. In this case, each column corresponds to a patient sample, while each row represents a gene. The color scale reflects expression intensity, where warmer colors indicate relatively higher expression levels and cooler colors indicate lower expression levels.

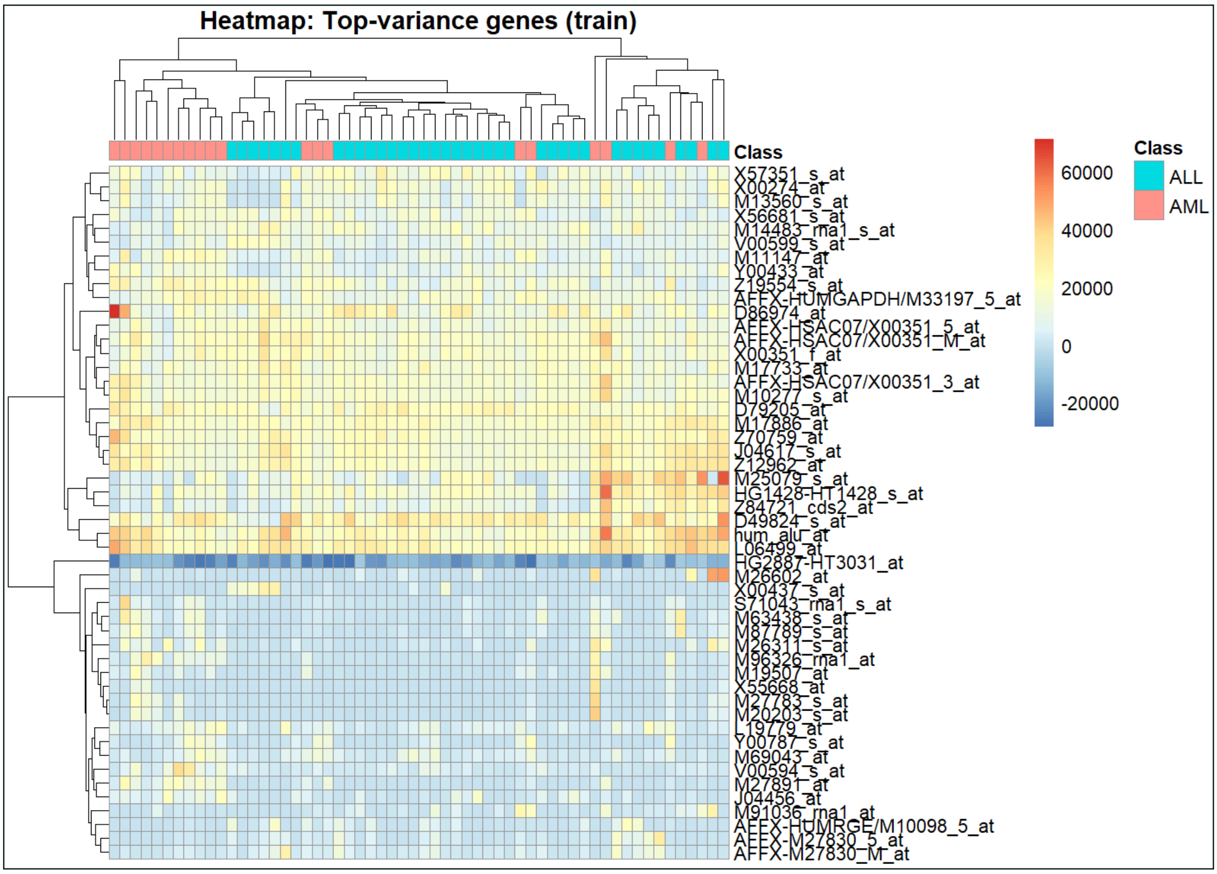

Figure 1: Hierarchical heatmap of the 50 genes with the highest variance in the training set

At the top of the heatmap, the annotation bar indicates the true class of each sample, distinguishing Acute Lymphoblastic Leukemia, or ALL, from Acute Myeloid Leukemia, or AML. One of the most important observations is the visible tendency of samples to cluster according to leukemia subtype. Although the separation is not mathematically perfect, we can clearly see that groups of AML samples tend to share similar expression patterns, while many ALL-samples cluster separately. This is an early indication that the dataset contains a biologically meaningful transcriptional signal capable of distinguishing the two diseases.

Another important aspect of the figure is the presence of large blocks of coordinated gene behavior. Certain groups of genes exhibit synchronized expression changes across multiple patients, forming structured expression patterns rather than random fluctuations. From a biological perspective, this may reflect shared regulatory pathways, lineage-specific transcriptional programs, or molecular mechanisms associated with lymphoid versus myeloid differentiation. Such coordinated behavior is particularly important in leukemia studies because these diseases are strongly driven by alterations in gene regulation and cellular differentiation processes.

The hierarchical clustering trees, shown both for genes and samples, also provide useful information about similarity relationships within the dataset. Samples located close to one another in the dendrogram tend to exhibit more similar global transcriptomic profiles. Likewise, genes clustered together often display correlated expression patterns and may participate in related biological pathways.

The heatmap allows visual identification of possible outliers or atypical samples. A few columns exhibit slightly different expression patterns compared to neighboring samples, which may reflect biological heterogeneity, technical variation, or patient-specific molecular characteristics. However, no extreme outlier appears dominant enough to distort the overall structure of the dataset.

This figure provides strong preliminary evidence that the ALL and AML classes are associated with distinct transcriptomic signatures. Therefore, the high predictive performance observed later in the machine learning pipeline is biologically plausible and not entirely unexpected. Nevertheless, because the sample size remains relatively limited, careful validation procedures are still necessary in order to avoid overestimating the generalizability of the model.

A visible tendency for separation of the samples according to class can be observed, suggesting that the ALL/AML signal is strong. In such cases, successful classification is not unexpected; however, caution is required due to the small sample size (small n).

Differential Expression Analysis with limma: Model, Testing and FDR

The limma framework applies linear regression independently for each gene. For each gene g and sample i:

$$x_{g,i} = \beta_{0,g} + \beta_{1,g} \cdot \mathbb{I}(y_i = \mathrm{AML}) + \varepsilon_{g,i}, \varepsilon_{g,i} \sim (0, \sigma_g^2)$$

where $\beta_{1,g}$ represents the mean expression difference between AML and ALL, as applied here on the expression intensities.

The limma method does not rely on a simple t-test, but instead uses a moderated t-statistic, where the variance is stabilized through Empirical Bayes shrinkage (Wesolowski et al., 2013):

$$t_g^{mod} = \frac{\hat{\beta}_{1,g}}{s_{g,post} \cdot \sqrt{v_g}}$$

where $s_{g,posts}$ is the posterior variance estimate and v_g is the relative variance of the estimator derived from the design matrix.

Because thousands of genes are tested simultaneously, multiple testing correction is required. Using the Benjamini–Hochberg False Discovery Rate (BH-FDR) procedure:
If $p(1) \le \cdots \le p(m)$ are the ordered p-values, then:

$$k = \max\left\{ i : p_{(i)} \le \frac{i}{m} q \right\}$$

and the hypotheses $H_{0,(1)}, \ldots, H_{0,(k)}$ are rejected, targeting an FDR level q (e.g., 0.05) (Storey, 2025).
After applying the variance filter (5,000 genes), the results showed:

- 734 genes significant at FDR≤0.05
- $\min(\text{adj.P.Val}) = 2.85 \times 10^{-10}$
- $\min(p) = 5.71 \times 10^{-14}$

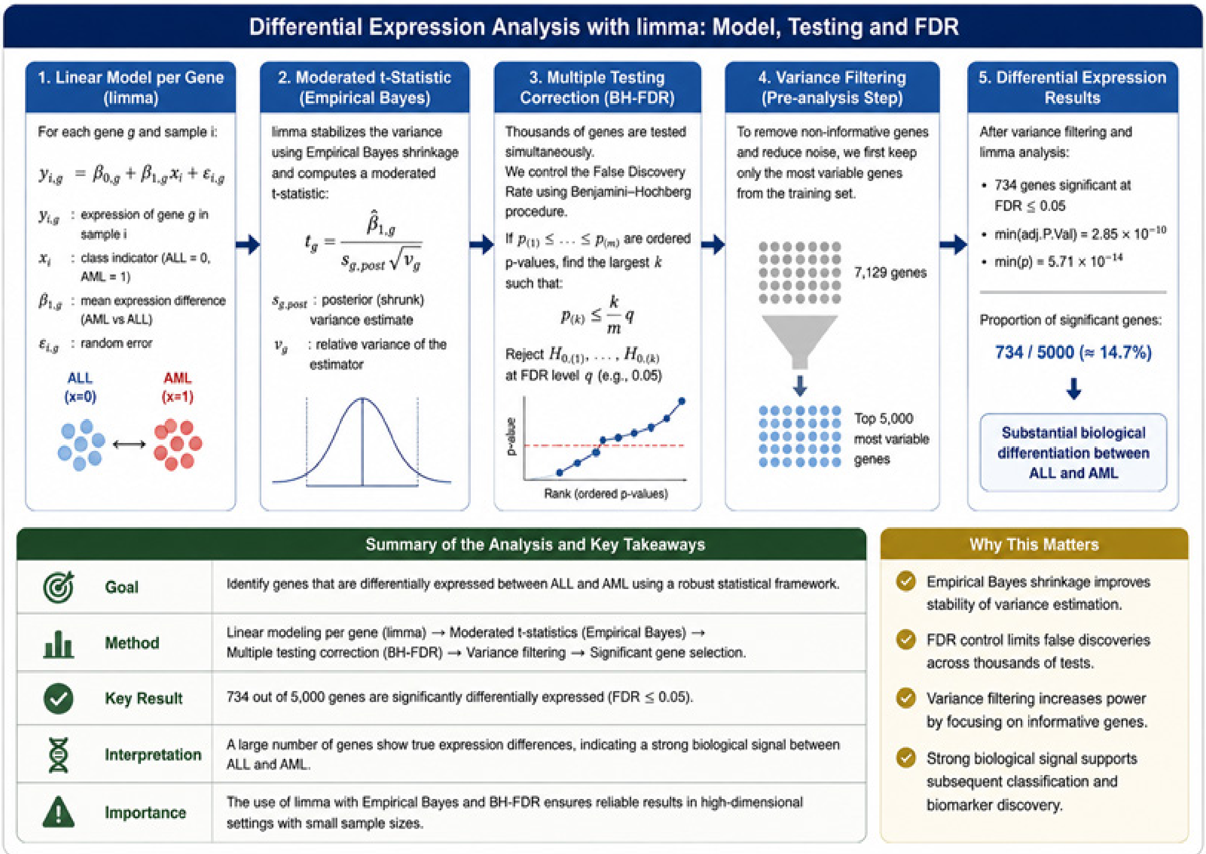

Figure 2: Differential Expression Analysis Pipeline Using limma and BH-FDR Correction for ALL vs AML Gene Expression Classification

The proportion of significant genes, 734/5000 (~14.7%), is substantial and consistent with strong biological differentiation between AML and ALL in this dataset, which has historically been known to produce a strong transcriptomic signal in leukemia classification studies (Cruz-Miranda et al., 2024).

Distribution of p-values and Volcano Plot

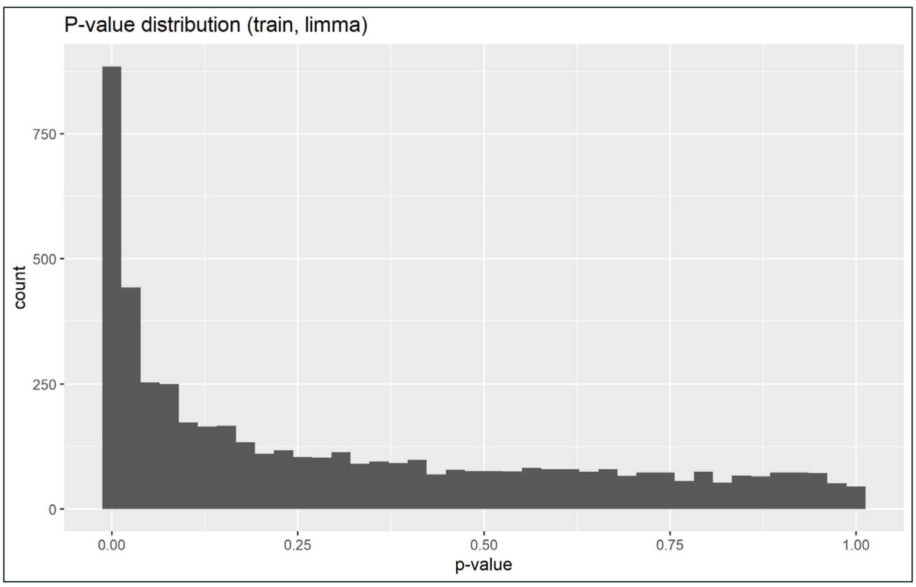

Figure 3: Histogram of p-values (limma) for all genes examined in the training set

The distribution shows a large concentration of p-values close to 0 and a more "flat" behavior across the remaining range. Qualitatively, this is consistent with a mixture of:

- genes with true differential expression (nonnull, small p-values),
- genes without differential expression (null, approximately uniform p-values).

The volcano plot visualizes effect size versus statistical significance:

$$y = -\log_{10}(p), x = \Delta_g$$

where here Δ_g is the estimated "logFC" returned by limma.

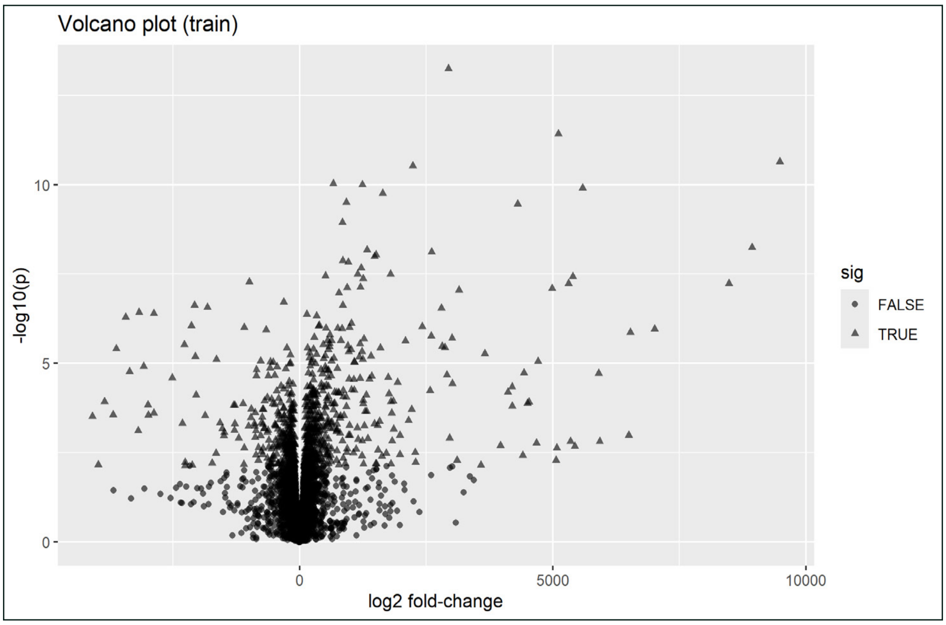

Figure 4: Volcano plot of the genes: the x-axis represents the log2 fold-change (logFC), while the y-axis represents −log10(p). Statistically significant genes (e.g., FDR ≤ 0.05) are highlighted.

In the code, the x-axis is labeled as "log2 fold-change"; however, based on the magnitude of the values (very large numbers), it appears that in practice the plot represents differences in mean expression intensity (on the microarray intensity scale) rather than true log2 fold-change values. This does not affect the statistical validity of the test, but it changes the interpretation of the "effect size": here it corresponds to differences in means on the original scale, rather than a ratio/log-ratio interpretation.

Gene Selection and Top Features Table

Based on the limma results obtained from the training set, K=200genes were selected. The selection rule was:

- ranking based on "adj.P.Val" (and subsequently on p-values),
- selection of the top K genes within the FDR threshold (which was sufficient in this case),
- fallback to |t| values if there were not enough statistically significant genes.

Typically:

$$\mathcal{G}_K = \{g : \mathrm{rank}(\mathrm{adjP}_g) \le K\}$$

Supervised PCA: Definition, Computation, and Interpretation

Let us assume that after feature selection, we retain the following expression matrix:

$$X \in \mathbb{R}^{p \times n}, p = 200, n = 58 \text{ (train)}$$

It is transformed into (samples × genes), and subsequently centered and scaled for each gene:

$$\tilde{x}_{i,g} = \frac{x_{i,g} - \mu_g}{\sigma_g}$$

PCA identifies the eigenvectors of the covariance matrix: $S = \frac{1}{n-1}\tilde{X}^{\top}\tilde{X}$

$$Sv_k = \lambda_k v_k$$

The PC scores for sample i are:

$$z_{i,k} = \tilde{x}_i^{\top} v_k$$

The proportion of explained variance of PC k is:

$$\mathrm{PVE}_k = \frac{\lambda_k}{\sum_j \lambda_j}$$

In the results::

- PC1 ≈ 0.4098 (≈ 41%)
- PC2 ≈ 0.0813
- PC3 ≈ 0.0666
- PC4 ≈ 0.0476
- PC5 ≈ 0.035

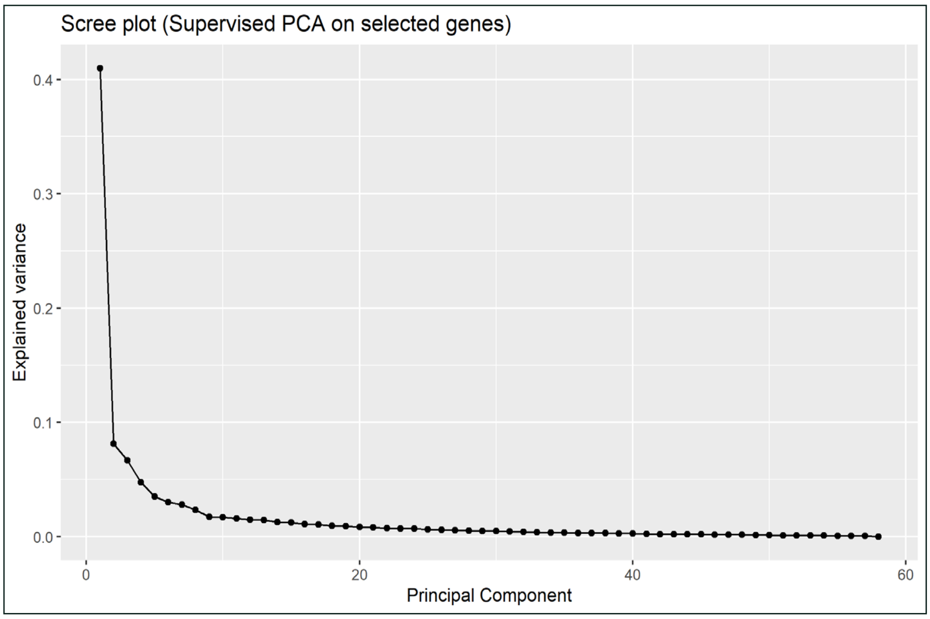

Figure 5: Scree plot of the proportion of explained variance per principal component, derived from PCA applied to the selected genes.

The very high proportion of variance explained (PVE) by PC1 suggests that a large part of the variability (within the subset of the "discriminative" genes) is concentrated in a single dimension—often an indication of strong class differentiation.

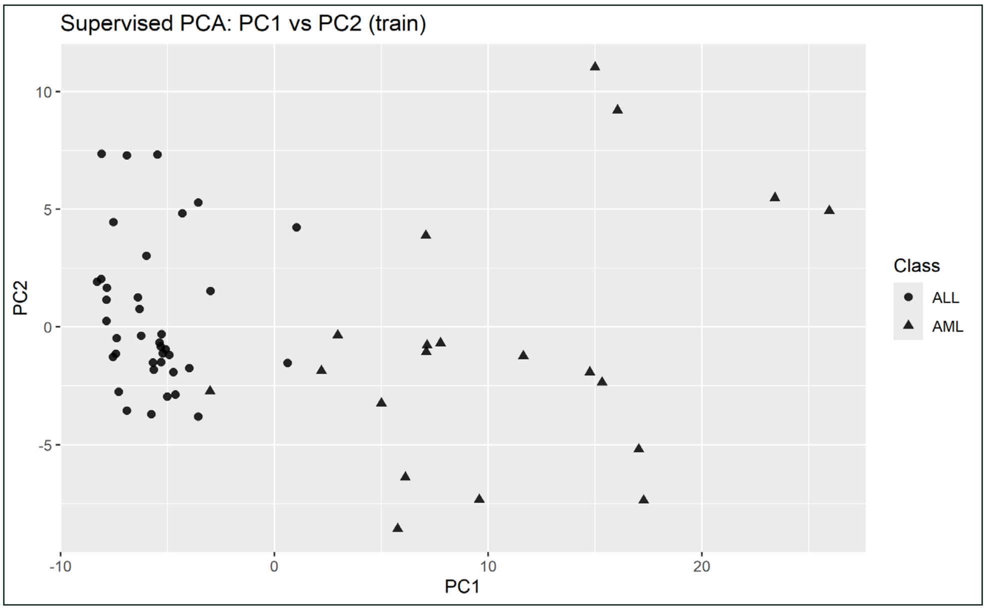

Figure 6: Scatter plot of the training samples in the space of the first two principal components (PC1, PC2) obtained from PCA applied to the selected (supervised) genes. Each point represents a sample, with shape/label according to class (ALL/AML)

In the training data, strong separation based on PC1 is observed. Indicatively, the mean value of PC1 for each

$\bar{z}_{ALL} \approx -5.64, \bar{z}_{AML} \approx 10.71$

This separation between class means largely explains the very high performance of the classifier: essentially, PC1 acts as an "almost linear separator."

Classification with Ridge Logistic Regression on the PCs

For each sample i, with feature vector z_i (here corresponding to the principal components), the logistic model is defined as:

$$\Pr(Y_i = 1 \mid z_i) = \pi_i = \frac{1}{1 + \exp(-(\beta_0 + \beta^T z_i))}$$

where Yi=1 corresponds to the AML class (positive class).
Ridge logistic regression estimates the parameters by solving:

$$\min_{\beta_0, \beta} \left[-\sum_{i=1}^{n} (y_i \log \pi_i + (1 - y_i)\log(1 - \pi_i)) + \lambda \parallel \beta \parallel_2^2 \right]$$

The parameter λ was selected through internal cross-validation (cv.glmnet, nfolds=5).

The use of ridge regression is particularly appropriate when multicollinearity/correlation exists among the PCs or when parameter stability is desired.

Using a threshold of 0.5:

$$\hat{y}_i = \begin{cases} AML, & \pi_i \geq 0.5 \\ ALL, & \pi_i < 0.5 \end{cases}$$

Simulation Results

We define:

- **TP**: AML correctly classified as AML
- **TN**: ALL correctly classified as ALL
- **FP**: ALL classified as AML
- **FN**: AML classified as ALL

In the test set, a perfect confusion matrix was obtained:

- **TN = 9, TP = 5, FP = 0, FN = 0**

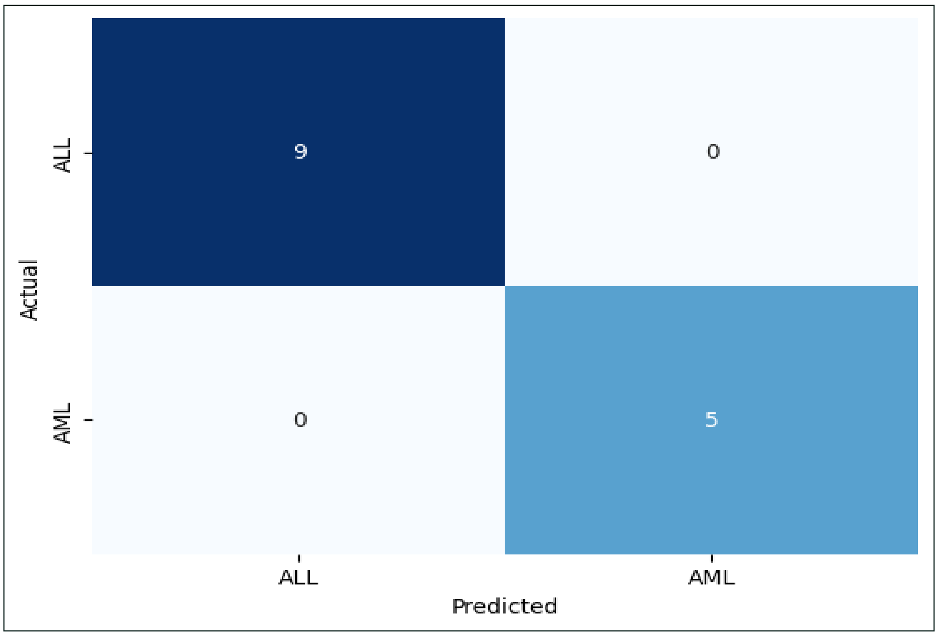

Figure 7: Confusion matrix on the test set (rows: true class, columns: predicted class), with absolute counts.

In addition, we define Accuracy, Sensitivity, Specificity, Precision, and F1-score as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Sensitivity (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity (TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Precision (PPV)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Having TP=5, TN=9, FP=0, FN=0:

- Accuracy = 1.0
- Sensitivity = 1.0
- Specificity = 1.0
- Precision = 1.0
- F1 = 1.0

The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for threshold t,

$$FPR(t) = \frac{FP(t)}{FP(t) + TN(t)}, TPR(t) = \frac{TP(t)}{TP(t) + FN(t)}$$

The AUC is defined as:

$$AUC = Pr(s(X^+) > s(X^-))$$

that is, the probability that a randomly selected positive sample has a higher score than a randomly selected negative sample.

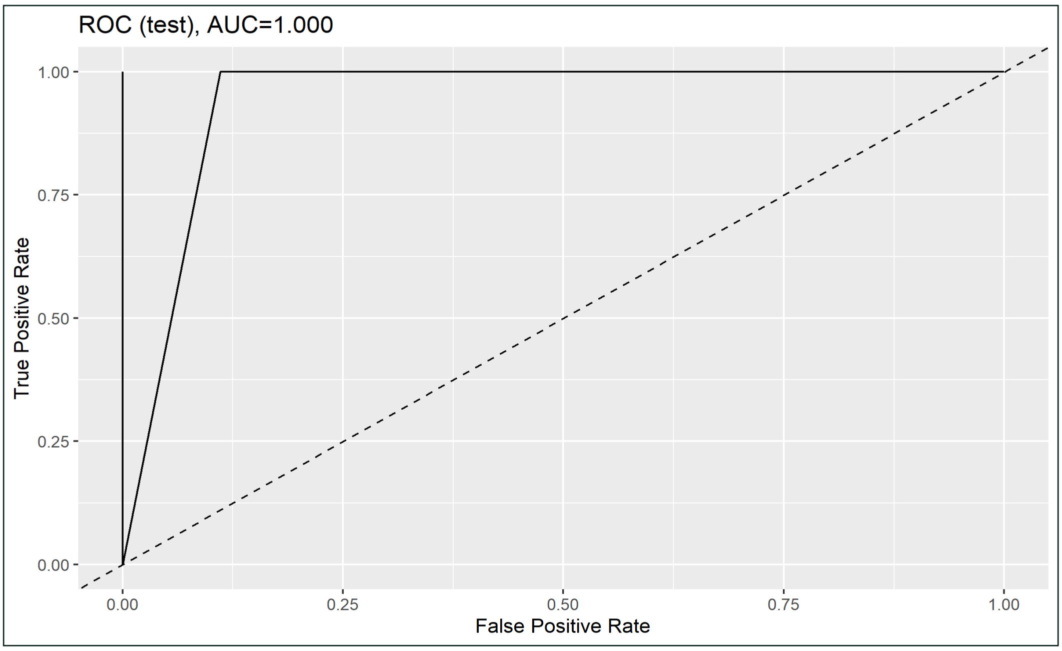

Figure 8: ROC curve on the test set for the classifier (based on the PCs), including the AUC value. The dashed diagonal line represents a random classifier.

As shown in the figure, a perfect AUC was obtained. In such a small test set (14 samples), this is feasible, especially when strong class separation exists. Nevertheless, it should not be interpreted as a "general" guarantee of performance on new clinical data—external validation is still required.

A 5-fold cross-validation (CV) procedure was applied on the training set (n=58n = 58n=58), with complete reconstruction of the pipeline within each fold:

- variance filtering only on the training fold,
- limma analysis only on the training fold,
- selection of 200 genes only on the training fold,
- PCA fitting only on the training fold,
- ridge logistic regression only on the training fold,
- prediction on the validation fold.

This procedure is critical for avoiding leakage, because otherwise the selected variables would "remember" information from the validation samples.
The results were:

- **Test AUC:** 1.0000
- **CV AUC (5-fold, no leakage):** 0.9829
- **Test Accuracy:** 1.0000

The CV AUC is slightly lower than the test AUC (=1), which is entirely reasonable: cross-validation estimates average performance across multiple splits and is generally more "strict." The value of 0.983 nevertheless remains extremely high and consistent with a strong biological signal.

To verify that the procedure does not produce a high AUC due to leakage or overfitting, a permutation test was applied:

- Random permutation of the labels $y \to y^{(b)}$
- Re-execution of the CV pipeline and computation of $AUC^{(b)}$
- Comparison of the observed performance with the null distribution under H_0

Typically:

$$y^{(b)} = \text{permute}(y), AUC^{(b)} = AUC(\text{pipeline}(X, y^{(b)}))$$

If a permutation p-value were to be computed:

$$p_{perm} = \frac{1 + \sum_{b=1}^{B} \mathbb{I}(AUC^{(b)} \geq AUC_{obs})}{B + 1}$$

With B=200, the following results were obtained:

- mean permuted CV AUC ≈ 0.497
- sd ≈ 0.088

This is exactly what would be expected under randomized labels (close to 0.5). Therefore, the high performance of the actual model is not an artificial consequence of the procedure, but rather reflects genuine information contained in the data.

Conclusion

The main conclusion of the analysis is that, in this specific problem (ALL vs AML), there is an exceptionally strong biological signal that becomes clearly evident already from the early stages of the pipeline. Limma, applied exclusively on the training set and after variance filtering (5,000 genes), identified a very large number of statistically significant genes (734/5000 with FDR ≤ 0.05), which is consistent with a systematic expression difference between the two classes. The distribution of p-values shows a strong concentration near zero, confirming that the observed signal does not correspond to random noise but rather to a large number of genuine expression differences. From a biomedical perspective, this finding further supports the established molecular heterogeneity between lymphoid and myeloid leukemias, reflecting distinct transcriptional programs, differentiation pathways, and oncogenic mechanisms associated with ALL and AML (Ribeiro et al., 2026).

At the same time, supervised PCA applied to the selected genes summarizes this differentiation into a small number of dimensions: PC1 explains approximately 41% of the variance and acts as the main axis of separation, while the PC1–PC2 plot geometrically demonstrates the clear separation of the samples. Thus, dimensionality reduction is not merely a computational convenience; it also constitutes strong and visually interpretable evidence that class membership is associated with a coherent expression pattern across a subset of genes. Clinically, this type of transcriptomic separation is highly relevant because accurate subtype identification is directly linked to therapeutic selection, prognosis estimation, and treatment stratification in leukemia management (Morcos-Sandino et al., 2025).

Following the above, the second conclusion is that the separation is largely almost linear in the space of the principal components; therefore, a simple linear model is sufficient and expected to achieve excellent performance. The PC1 vs PC2 plot shows that the AML samples cluster in a region clearly separated from the ALL samples mainly along PC1, which explains why ridge logistic regression (with regularization) achieves very high performance without requiring a nonlinear classifier. The choice of ridge regression is methodologically appropriate because it stabilizes parameter estimation in a high-dimensional and limited-sample environment, avoiding extreme coefficients and overfitting. In addition, the use of cross-validated selection of the parameter λ (through cv.glmnet) provides an internally controlled adjustment of model complexity. The "perfect" performance observed on the test set (Accuracy = 1, AUC = 1) does not appear here as a suspicious finding by itself, but rather as a natural consequence of a dataset with a strong signal and clear class separation, especially when the feature engineering process (limma → gene selection → supervised PCA) explicitly aligns the feature space with the class discrimination structure. From a translational medicine perspective, these findings highlight the growing role of interpretable artificial intelligence systems in supporting hematological diagnostics and molecular decision-support systems (Sartori et al., 2025).

The third and most critical conclusion concerns the reliability of the evaluation procedure and its limitations. The cross-validation process was implemented without information leakage (no leakage), since variance filtering, limma/gene selection, and PCA were all applied within each fold exclusively on the training portion of that fold. The very high performance obtained through cross-validation within the training set (CV AUC ≈ 0.983) indicates that the model generalizes consistently to "unseen" subsets of the training data and does not rely on random idiosyncrasies. Even more importantly, the permutation sanity check (200 random label permutations) produced a mean permuted CV AUC ≈ 0.497 with sd ≈ 0.088, that is, close to 0.5 as expected under random labeling; this constitutes strong evidence that the pipeline does not "leak" information.

Nevertheless, the main limitation is the small test set (n=14), which makes the observed Accuracy/AUC = 1 potentially fragile under a different sampling realization. In addition, the interpretation of logFC requires caution, because it does not necessarily correspond to a biological "log2 fold-change," but rather to differences in means on the intensity scale (unless an appropriate log-transform has previously been applied). Another important limitation is that the analysis was based on a single historical microarray dataset; therefore, external validation using independent cohorts and modern RNA-seq datasets would be necessary before considering clinical implementation. Furthermore, although the model demonstrates excellent discrimination performance, transcriptomic biomarkers in hematological malignancies may be influenced by biological heterogeneity, treatment status, sample preparation procedures, and population-specific genomic variation (Capelletti et al., 2025).

Overall, however, the pipeline limma → gene selection → supervised PCA → ridge logistic regression proves to be highly effective, interpretable, and methodologically consistent for high-dimensional problems characterized by a strong biological signal. The study demonstrates how the integration of statistical genomics and explainable machine learning can contribute to precision oncology by improving disease subtype classification while maintaining methodological transparency and reproducibility. More broadly, such approaches may support the future development of AI-assisted diagnostic frameworks in hematology and personalized medicine, where accurate molecular characterization is increasingly essential for optimized patient care and targeted therapeutic interventions (Santoro et al., 2024).

References

1. Ali A, Khan Z, Aldahmani S (2025) Optimized feature selection in high-dimensional gene expression data using weighted differential gene expression analysis. Applied Soft Computing 180: 113329.
2. Apicella A, Isgrò F, Prevete R (2025) Don't push the button! Exploring data leakage risks in machine learning and transfer learning. Artif Intell Rev 58: 339.
3. Araújo R, Ramalhete L, Viegas A, Von Rekowski CP, Fonseca, et al. (2024) Simplifying Data Analysis in Biomedical Research: An Automated, User-Friendly Tool. Methods and Protocols 7: 36.
4. Capelletti MM, Montini O, Ruini E, Tettamanti S, Savino AM, et al. (2025). Unlocking the Heterogeneity in Acute Leukaemia: Dissection of Clonal Architecture and Metabolic Properties for Clinical Interventions. International Journal of Molecular Sciences 26: 45.
5. Cruz-Miranda GM, Olarte-Carrillo I, Bárcenas-López DA, Martínez-Tovar A, Ramírez-Bello, et al. (2024) Transcriptome Analysis in Mexican Adults with Acute Lymphoblastic Leukemia. International Journal of Molecular Sciences 25: 1750.
6. Jha M, Hasija Y (2026) Development and validation of AI-driven multi-omics language models for cancer genomics: A comprehensive review. Computational Biology and Chemistry 120: 108662.
7. Lin W, Feng R, Li H (2015) Regularization Methods for High-Dimensional Instrumental Variables Regression with an Application to Genetical Genomics. Journal of the American Statistical Association 110: 270-288.
8. Morcos-Sandino M, Quezada-Ramírez SI, Gómez-De León A (2025) Advances in the Treatment of Acute Myeloid Leukemia: Implications for Low- and Middle-Income Countries. Biomedicines 13: 1221.
9. Ribeiro ÂV, Alves GM, Rodrigues GSB, Silva CDCM, Barreto MA, et al. (2026) Applications of Artificial Intelligence in the Diagnosis of Acute Promyelocytic Leukemia: A Bibliographic Review. Proceedings 137: 65.
10. Santoro N, Salutari P, Di Ianni M, Marra A (2024) Precision Medicine Approaches in Acute Myeloid

Leukemia with Adverse Genetics. International Journal of Molecular Sciences 25: 4259.

11. Sartori F, Codicè F, Caranzano I, Rollo C, Birolo G, et al. (2025) A Comprehensive Review of Deep Learning Applications with Multi-Omics Data in Cancer Research. Genes, 16(6), 648.
12. Storey JD (2025) False Discovery Rate. In: Lovric, M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg.
13. Tsega A, Mullualem D (2026) Machine learning for multi-omics data integration in crop improvement: a systematic review. BMC Bioinformatics 27: 81.
14. Wang J, Zhang Z, Wang Y (2025) Utilizing Feature Selection Techniques for AI-Driven Tumor Subtype Classification: Enhancing Precision in Cancer Diagnostics. Biomolecules 15: 81.
15. Wesolowski S, Birtwistle MR, Rempala GA (2013) A Comparison of Methods for RNA-Seq Differential Expression Analysis and a New Empirical Bayes Approach. Biosensors 3: 238-258.