DOI: https://doi.org

International Journal of Advanced and Innovative Research
Journal homepage :https://scholarclub.org/index.php/IJAIR/login

Convergent Feature Selection and Predictive Modeling Identify AIM2 as a Potential Molecular Diagnostic Marker in Acute Myeloid Leukemia

Shehzad Khalil(Coresponding Author)
shehzadkhalil911@gmail.com
Institute of Biotechnology and Genetic Engineering, The University of Agriculture, Peshawar, Khyber Pakhtunkhwa, Pakistan
Fayaz Ahmad Khan
fayazahmadkhan277@gmail.com
Department of Biotechnology, Abdul Wali Khan University, Mardan 23200, Khyber Pakhtunkhwa, Pakistan
Ahmed Mujtaba Veesar
ahmedabroahmedabro6@gmail.com
Institute of Biotechnology and Genetic Engineering, The University of Sindh, Jamshoro, Sindh, Pakistan
Zain Uddin
uzain9372@gmail.com
Department of Biotechnology, Balochistan University of Information Technology, Engineering and Management Sciences, Pakistan
Halaar Hussain
dhalaar@gmail.com
Institute of Biotechnology and Genetic Engineering, The University of Sindh, Jamshoro, Sindh, Pakistan
Abubakar Basiru
abubakar.basiru@fuhsi.edu.ng
Department of Medical Laboratory Science, Federal University of Health Sciences, Ila Orangun, Osun State, Nigeria
Abbas Ahmad
Department of Biotechnology, Abdul Wali Khan University, Mardan 23200, Khyber Pakhtunkhwa, Pakistan

ARTICLE INFO: **ABSTRACT**

Received:
17 06 2026
Revised:
05 07 2026
Accepted:
25 07 2026

Keywords:
Acute myeloid leukemia,AIM2;,Biomarker; Transcriptomics,Machine learning; Random Forest, SVM-RFE, Inflammasome,RNA-sequencing, Differential expression

Background: *Acute myeloid leukemia (AML) is an aggressive hematologic malignancy that occurs by clonal expansion of myeloid progenitor cells, is highly molecularly heterogeneous, and has a poor prognosis and low precision of diagnosis. Defining strong transcriptomic signatures that can differentiate AML from normal hematopoietic progenitors would be critical to increase the accuracy of diagnosis and the therapeutic stratification.*
Methods: *RNA-sequencing data were obtained from GEO (accession number GSE235686) for 15 AML patients and 2 CD34+ healthy hematopoietic stem/progenitor cells (HSPC) controls. The limma-voom framework was then used for differential expression analysis after data preprocessing and data quality filtering. A composite scoring system using multiple metrics was built to rank candidate biomarkers. The functional enrichment analysis was performed with clusterProfiler and ShinyGO v0.8. The Wilcoxon rank-sum test with AUC filtering, Random Forest feature importance, and Support Vector Machine with Recursive Feature Elimination (SVM-RFE) were used as three complementary machine learning algorithms to determine a final biomarker panel. Leave one out cross validation (LOOCV) was used to assess the classification performance on six machine learning models.*
Results: *There were 857 significantly differentially expressed genes (DEGs) identified ($|\log_2 FC| > 1$, FDR < 0.05); 631 of these (73.6%) were up-regulated in AML. Functional enrichment identified key pathways of immune response, cytokine*

signaling and DNA replication as dysregulated. A composite score was created that favoured 30 candidate biomarkers that were all up-regulated in patients with AML with complete consistency. A multi-algorithm feature selection led to a set of 13 final biomarkers (selected by two or more algorithms). The only gene that was selected by all three algorithms (AUC = 1.000) was Absent in Melanoma 2 (AIM2). Random Forest classification had a perfect LOOCV performance in terms of Accuracy (1.000), AUC (1.000), Sensitivity (1.000) and Specificity (1.000) compared with Logistic Regression (AUC = 0.733).

***Conclusions**: Incorporation of transcriptomic profiling and multi-algorithm machine learning is an effective strategy for discovering strong AML biomarkers. AIM2 is the most consistently reported candidate as it has been known to play a role in innate immune signaling through the inflammasome, and has not been found in normal CD34+ HSPCs. These results may be used to establish that AIM2 is a diagnostic and therapeutic target in AML which could be tested in larger cohorts in the future.*

INTRODUCTION

Acute myeloid leukemia (AML) is a fast-growing accumulation of immature cells in the blood and bone marrow that are called myeloid blasts, and is one of the most frequent and deadly hematological malignancies in adults. AML is a significant clinical and hematological problem with an estimated incidence of 4.3 per 100,000 in the USA and an overall 5-year survival of around 30–40% [1]. The molecular basis of the disease is extremely heterogeneous, with a range of chromosomal abnormalities, somatic mutations, and epigenetic changes all contributing to normal haematopoietic differentiation being disrupted and to malignant transformation [2]. The current diagnosis of AML is based mainly on the morphological examination of bone marrow aspirates, cytogenetic karyotyping, and a few molecular markers such as FLT3 mutations, NPM1 mutations, CEBPA mutations and IDH1/2 mutations [3]. The 2022 World Health Organization (WHO) classification and the European LeukemiaNet (ELN) guidelines include the introduction of molecular profiling in risk stratification, but these were not sufficiently sensitive to identify all the relevant AML disease subtypes, nor do they fully reflect the transcriptomic complexity of AML [4].This means that many AML patients suffer from relapse after receiving the traditional induction chemotherapy; thus, the need for discovering new, strong molecular markers for diagnostic accuracy and development of new drugs is still urgent. The RNA sequencing (RNA-seq) technique has revolutionized the way in which the transcriptomic profile of a hematologic malignancy can be studied by allowing the genome-wide quantification of gene expression with high sensitivity and specificity [5]. Comparing expression levels of malignant and normal hematopoietic precursors can reveal pathways and candidate biomarkers that conventional approaches are not able to detect. A comparison is especially interesting between AML cells and CD34+ hematopoietic stem/progenitor cells (HSPCs), because AML blasts are believed to originate from, and share immunophenotypic characteristics with, CD34+ progenitors [6]. In the field of oncology, machine learning (ML) algorithms have emerged as a valuable tool for discovering biomarkers, patient stratification and prediction of outcomes, as they have been increasingly used in this area [7]. High dimensional genomic data features are well suited to be handled by ensemble methods like random forest or discriminative classifiers like support vector machines (SVMs), which can discover non-linear relationships between molecular features [8]. But in rare malignancies, small clinical cohorts pose a major problem in training and validating models. These limitations can be somewhat overcome by using composite scoring, multi-algorithm feature selection and leave-one-out cross-validation (LOOCV), which helps to prevent overfitting and to retain features which are consistent when analysed with several algorithms [9]. AIM2 (Absent in Melanoma 2) is a PYHIN family cytosolic innate immune sensor which detects double-stranded DNA (dsDNA) in the cytoplasm, initiates the formation of the AIM2 inflammasome, which in turn leads to the activation of caspase-1 and the maturation of interleukin-1β (IL-1β) and IL-18 [10]. Although AIM2 has been extensively studied in the context of bacterial and viral infections, and autoimmunity, recent evidence has suggested a role for AIM2 in the activation of the inflammasome in hematologic malignancies such as leukemia [11]. However, the function of AIM2 in AML is still poorly understood. In this study, we used a novel multi-metric composite biomarker scoring framework, multi-algorithm machine learning feature selection, and complete transcriptomic profiling of RNA-seq data in AML patients and healthy CD34+ HSPC controls (GEO accession: GSE235686) to systematically identify and prioritize candidate AML biomarkers. An integrated approach reveals that AIM2 is the most consistently prioritized candidate, is selected by all three independent feature selection methods and has perfect individual diagnostic performance, making it a potential novel biomarker for AML diagnosis.

2. **Materials and Methods**

Figure 1 illustrates the overall workflow of the study methodology used for biomarker identification and diagnostic evaluation. The pipeline begins with data source collection and preprocessing, followed by differential expression analysis and multi-metric composite biomarker scoring. Functional enrichment analysis was performed to investigate the biological significance of the identified genes. Subsequently, machine learning-based feature selection approaches, including the Wilcoxon rank-sum test, Random Forest feature importance, and SVM with recursive feature elimination, were applied to identify the most informative biomarkers. Finally, the selected biomarkers were integrated into a final diagnostic panel, and their performance was evaluated using diagnostic accuracy metrics.

Figure 1: Workflow of the biomarker identification and diagnostic evaluation pipeline.

2.1 **Data Source and Study Design**

RNA-sequencing data for 15 AML patients and 2 healthy controls of CD34+ hematopoietic stem/progenitor cells (HSPC) were downloaded from the Gene Expression Omnibus (GEO) database (Accession No. GSE235686) [12]. The data included 20,000 or so protein-coding genes for 17 samples of processed expression values (TPM/1000 normalized). Bone marrow aspirates were used to obtain all AML samples, and CD34+ cells from healthy donors were used as healthy control. The data quality, availability of matching control samples and clinical relevance of the comparison group were used to select this dataset.

2.2 **Data Preprocessing**

All analyses were conducted using the R software (version 4.6.0). The expression matrix was imported and duplicate gene symbols were cleaned using the function make.unique [13]. The sample identifiers were standardized by removing the dots and replace them with hyphens, so that they can be compatible with the analytical tools used. All of the expression values were shifted by adding a pseudocount of 0.001 and then $\log_2$ transformed. After transformation, genes with no variation between any of the 17 samples were discarded. Furthermore, genes with expression below the 25th percentile (lowly expressed) were removed to lower background noise and increase the statistical power in subsequent analyses [14]. After quality control, 27199 genes were retained for further analysis.

2.3 Differential Expression Analysis

We performed differential expression analysis between AML patients (n = 15) and healthy controls (CD34+ n = 2) using the limma-voom framework that employs empirical Bayes moderation [25, 30]. Condition was a binary variable (AML vs. Healthy) in the design matrix [15]. Due to the small number of healthy replicates, the trend = TRUE parameter was used for empirical Bayes moderation to borrow information across genes to stabilize the variance and improve the statistical estimates in small sample sizes. Genes with an absolute $\log_2$ fold change $|\log_2 FC| > 1$ (two-fold difference) and Benjamini-Hochberg adjusted p-value (false discovery rate, FDR) < 0.05 were considered as differentially expressed [16].

2.4 Multi-Metric Composite Biomarker Scoring

To address the limitation of small healthy control numbers (n = 2) and to prioritize robust biomarker candidates beyond standard statistical thresholds, a multi-metric composite scoring approach was developed [17]. For each significantly differentially expressed gene, five metrics were calculated:

(i) **Consistency score (C**): The proportion of AML patients exhibiting expression levels above the maximum observed in healthy controls, directly quantifying the gene's ability to separate disease from normal samples.

(ii) **Fold separation** (S): The ratio of median AML expression to maximum healthy expression, measuring the magnitude of expression separation.

(iii) **Coefficient of variation** (CV): Standard deviation divided by mean expression across AML patients; lower values indicate more consistent intra-group expression.

(iv) **Statistical significance**: Quantified as -log10(adjusted p-value).

(v) Healthy expression penalty (H) and direction bonus (D): A tiered penalty was applied to genes with elevated baseline expression in healthy controls (H = 1.0 if healthy max < 0.01; H = 0.8 if < 0.05; H = 0.5 if < 0.1; H = 0.2 otherwise). An up-regulation bonus (D = 1.2) was applied to up-regulated genes; down-regulated genes received D = 0.8.

Each continuous metric was normalized to a [0, 1] scale using min-max normalization as:

The composite biomarker score was computed as:

Score=0.40*C_norm+0.10*|〖log〗_2 FC|_norm+ 0.5*S_norm+0.10*(-log_10 (Padj))_norm+0.10*(1-CV_norm)* H* D

The consistency score received the highest weighting (40%) to prioritize genes with robust discriminatory ability despite the small reference group. Genes were ranked in descending order by composite score, and the top 30 candidates were selected for machine learning analysis.

2.5 **Functional Enrichment Analysis**

The clusterProfiler package (version 4.20.0) was used in R (version 4.3.0) [18] to perform Gene Ontology (GO) [19]and KEGG pathway enrichment analysis [19] on the set of all significant DEGs with org.Hs.eg.db (version 3.17.0) as the annotation database [20]. Each of the three GO ontologies (Biological Process, Cellular Component, Molecular Function) was tested separately with a hypergeometric test and a Benjamini-Hochberg FDR correction. The terms that were found to be significantly enriched were those with adjusted p-value less than 0.05 (GO terms and KEGG pathways). The gene symbols were translated to Entrez IDs for KEGG analysis. Dot plots and bar plots were used to display the results [21]. The top 30 prioritized biomarkers were input into a web-based complementary functional enrichment analysis tool called ShinyGO v0.8 [22]. Analysis was performed with use of Homo sapiens as organism, GO and KEGG databases, FDR cutoff of 0.05 and pathway size range of genes from 2 to 2000 were used.

2.6 **Machine Learning Feature Selection**

To achieve this, three different complementary feature selection methods were applied to the top 50 genes with the highest composite-scored candidate gene expression (Z-score normalized) as features [23].

2.6.1 **Wilcoxon Rank-Sum Test**

For each of the 50 candidate genes, the Wilcoxon rank-sum test was used to compare expression distributions among AML (n = 15) and healthy (n = 2) samples. Selection criteria were a Benjamini-Hochberg adjusted p-value threshold of 0.2 (relaxed because of the small sample size) and AUC > 0.7 [24]. For the purpose of ranking the pseudo-coefficients were calculated as $\log_2(AUC/(1 - AUC) + 0.1)$.

2.6.2 **Random Forest Feature Importance**

The expression matrix is scaled and a Random Forest classifier with 500 decision trees is trained. The importance of each feature was calculated as Mean Decrease Gini, which represents the contribution made by each gene in terms of node purity [25]. The random forest feature set is the top 8 genes by Mean Decrease Gini.

2.6.3 **SVM with Recursive Feature Elimination**

Linear kernel was used for SVM-RFE, which was also done with LOOCV [26]. Since the analysis was computationally complex, only the top 30 Wilcoxon-selected genes were considered. The optimal feature subset of size 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 was chosen to maximize the accuracy of LOOCV, which are a subset of the parameters [27].

2.7 **Final Biomarker Panel and Diagnostic Performance Evaluation**

The final biomarkers panel was determined to be the overlapping genes of a gene selected by at least two of three feature selection approaches. Individual diagnostic performance was evaluated by ROC curve analysis by using the package pROC [28] with 95% CI estimation by DeLong method. The six classification algorithms were then tested on the final biomarker panel using LOOCV, all of which were implemented through the caret package (version 7.0.1) [24]: Logistic Regression, Random Forest (500 trees) [29], SVM-Linear [30], SVM-Radial [31], Naive Bayes [32], and Gradient Boosting [33] (150 iterations, interaction depth = 3, shrinkage = 0.1). Accuracies, AUC, sensitivity, specificity, negative predictive values (NPVs) and positive predictive values (PPVs) were used as performance metrics.

3. **RESULTS**

3.1 **Data Processing and Quality Control**

The RNA-seq data set consisted of 17 samples, 15 AML patients and 2 normal CD34+ HSPC controls. There were 60,483 raw expression values in the matrix. After QC, 27,199 genes (44.9% of the original transcripts) were selected for downstream analysis following removal of zero variance genes and low expression filtering (bottom 25th percentile).

3.2 **Principal Component Analysis**

Principal component analysis (PCA) of the filtered expression matrix showed good separation between AML patients and their healthy controls (CD34+) along the first principal component (PC1) that accounted for 18.8% of the total variance. PC2 explained 11.1% of the variance. The two healthy control samples were grouped very close together, indicating high reproducibility. The molecular heterogeneity of AML caused moderate dispersion along PC2. All 17 samples were included in subsequent analyses as there was no evidence of outlier samples (Figure 2a).

3.3 **Differential Expression Analysis**

We compared gene expression in AML to healthy CD34+ HSPCs by differential expression analysis, revealing a list of 857 genes that were significantly dysregulated ($|\log_2 FC| > 1$, FDR < 0.05). 631 (73.6%) genes were up-regulated and 226 (26.4%) were down-regulated in AML. The tendency toward transcriptional up-regulation is predominant because most genes are activated instead of suppressed in leukemic blasts compared to normal hematopoietic progenitors. A volcano plot has been used to show the distribution of effect sizes, which is consistent with the substantial transcriptomic reprogramming in AML (Figure 2b).

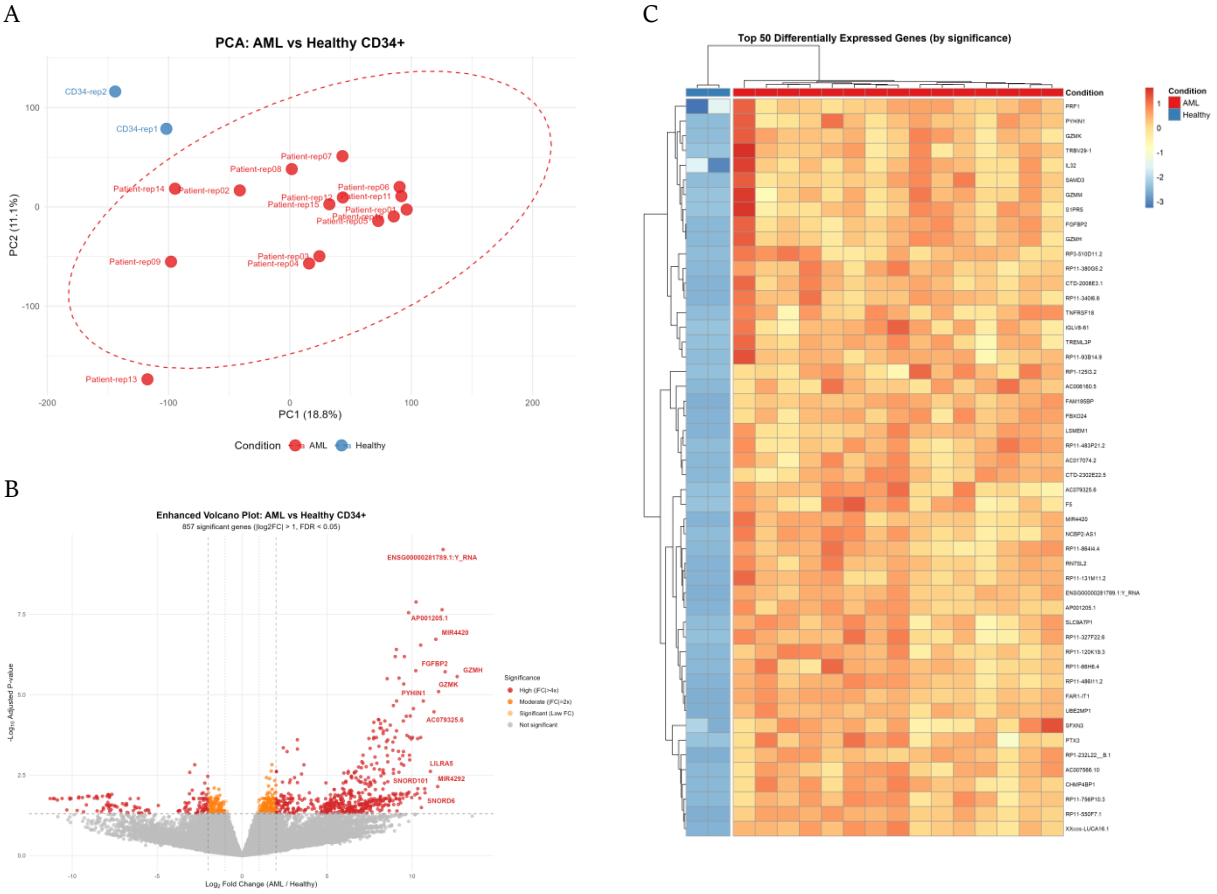


Figure 2: Differential expression landscape of AML versus healthy CD34+ controls. (a) PCA plot showing clear separation between AML patient samples (red) and healthy controls (blue) based on the first two principal components. (b) Enhanced volcano plot displaying 857 significantly up- and down-regulated genes (log_2FC > 1, FDR < 0.05$) in AML compared to healthy controls, with top significant genes labeled. (c) Bidirectional hierarchical clustering heatmap showing the expression profiles of the top 50 most significant differentially expressed genes across all samples.

3.4 Functional Enrichment Analysis of Differentially Expressed Genes

Gene Ontology Enrichment

The GO enrichment analysis of 857 significant DEGs showed that the functional patterns of the DEGs were different in both directions. Immune response pathways were highly overrepresented in the up-regulated genes (both "immune response-regulating cell surface receptor signaling pathway" and "immune response-activating cell surface receptor signaling pathway" were highly overrepresented, with adjusted $p < 0.002$) (Figure 3a-c). These results suggest that AML cells have an abnormal activation of surface receptor-mediated immune signaling pathways. Down-regulated genes were enriched in pathways associated with DNA replication and RNA processing such as "DNA replication" (adjusted $p < 0.01$), "rRNA processing" (adjusted $p < 0.01$) and "ribosome biogenesis" (adjusted $p < 0.01$). Importantly, the suppression of these fundamental cellular processes in AML are also consistent with the overall literature, which indicates that key pathways controlling ribosome biosynthesis and DNA synthesis are deregulated as a result of oncogenic transformation and the different cell-of-origin comparisons [28].

KEGG Pathway Enrichment

A KEGG pathway analysis revealed the presence of 23 significantly enriched pathways. The most enriched were "Primary immunodeficiency" (adjusted p < 0.001), "Chemokine signaling pathway" (adjusted p < 0.001; 12 genes), "Viral protein interaction with cytokine and cytokine receptor" (adjusted p < 0.001; 16 genes), and "Natural killer cell mediated cytotoxicity" (adjusted p < 0.001; 8 genes). Other significantly enriched pathways were "Cytokine-cytokine receptor interaction" (12 genes) and "T cell receptor signaling pathway" (16 genes), underscoring widespread immune signaling pathway abnormalities and cytokine pathways in AML (Figure 3d).

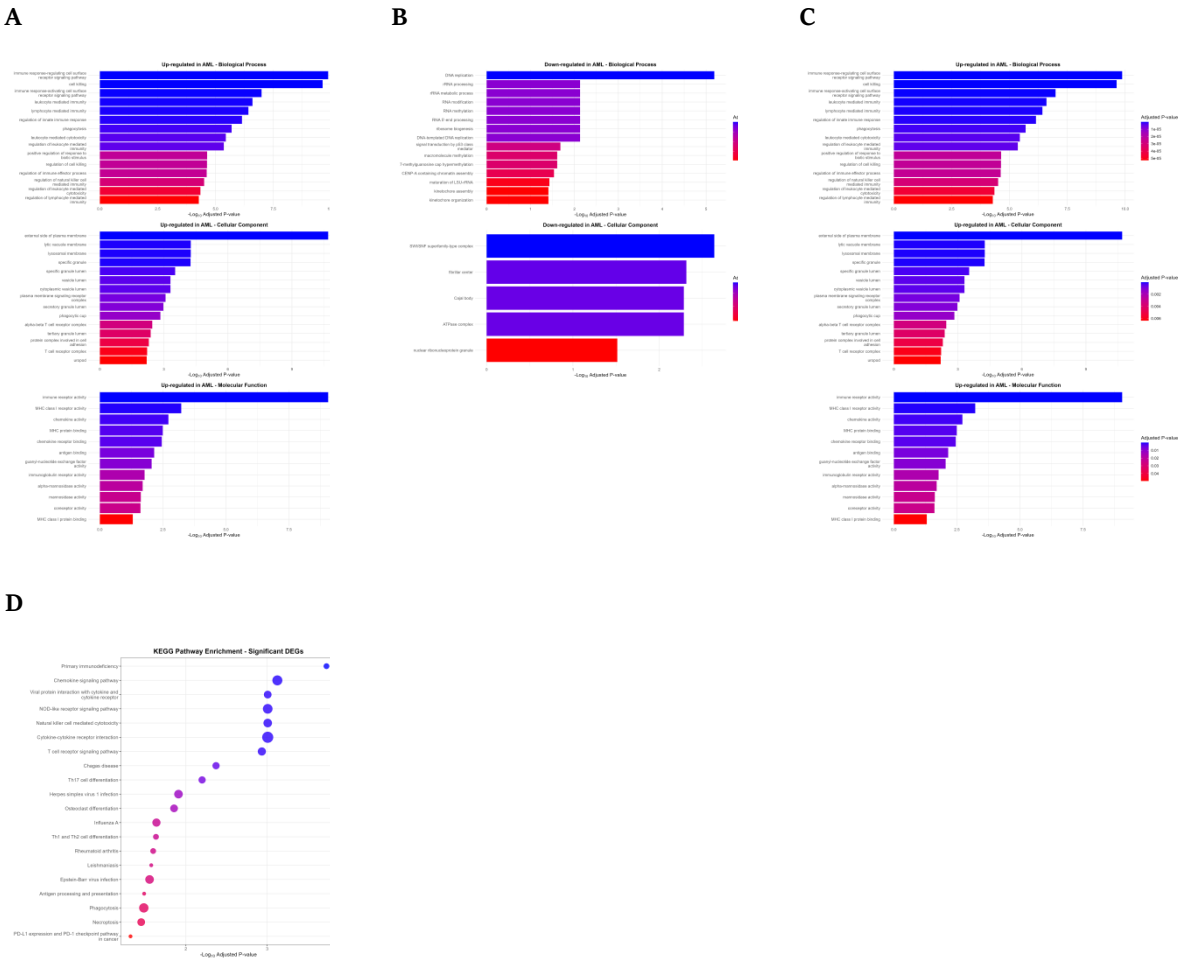


Figure 3: Functional enrichment and pathway analysis of differentially expressed genes. (a) Gene Ontology (GO) enrichment analysis for significantly up-regulated genes in AML, categorized by Biological Process (top), Cellular Component (middle), and Molecular Function (bottom). (b) GO enrichment analysis for significantly down-regulated genes across the same three categories. (c) Overarching GO enrichment analysis pooling all significant differentially expressed genes (DEGs). For (a), (b), and (c), the x-axis indicates statistical significance (-log_10 adjusted p-value), with bars color-coded by significance level. (d) KEGG pathway enrichment dot plot of significant DEGs, where dot size corresponds to gene count and color denotes the adjusted p-value.

3.5 Functional Enrichment of Top 30 Biomarker Candidates

The top 30 biomarker candidates were analyzed with ShinyGO, showing enrichment of innate immune pathways in Biological Process analysis (fold enrichment > 50), such as "activation of innate immune response," "cellular response to cytokine stimulus," "cellular response to tumor necrosis factor," and "positive regulation of immune system process. In particular, Cellular Component analysis revealed high enrichment of the "AIM2 inflammasome complex" (fold enrichment = 346.67), which points to the direct involvement of AIM2 and its interacting partners in the candidate set of biomarkers. Molecular Function analysis showed an enrichment of serine-type endopeptidase activity, peptidase activity and death receptor binding. To confirm the biological relevance of the AIM2-inflammasome signaling in the identified biomarker panel, KEGG analysis showed the enrichment of the pathway "Cytosolic DNA-sensing pathway" (fold enrichment = 20.5) (Figure 4a-c).

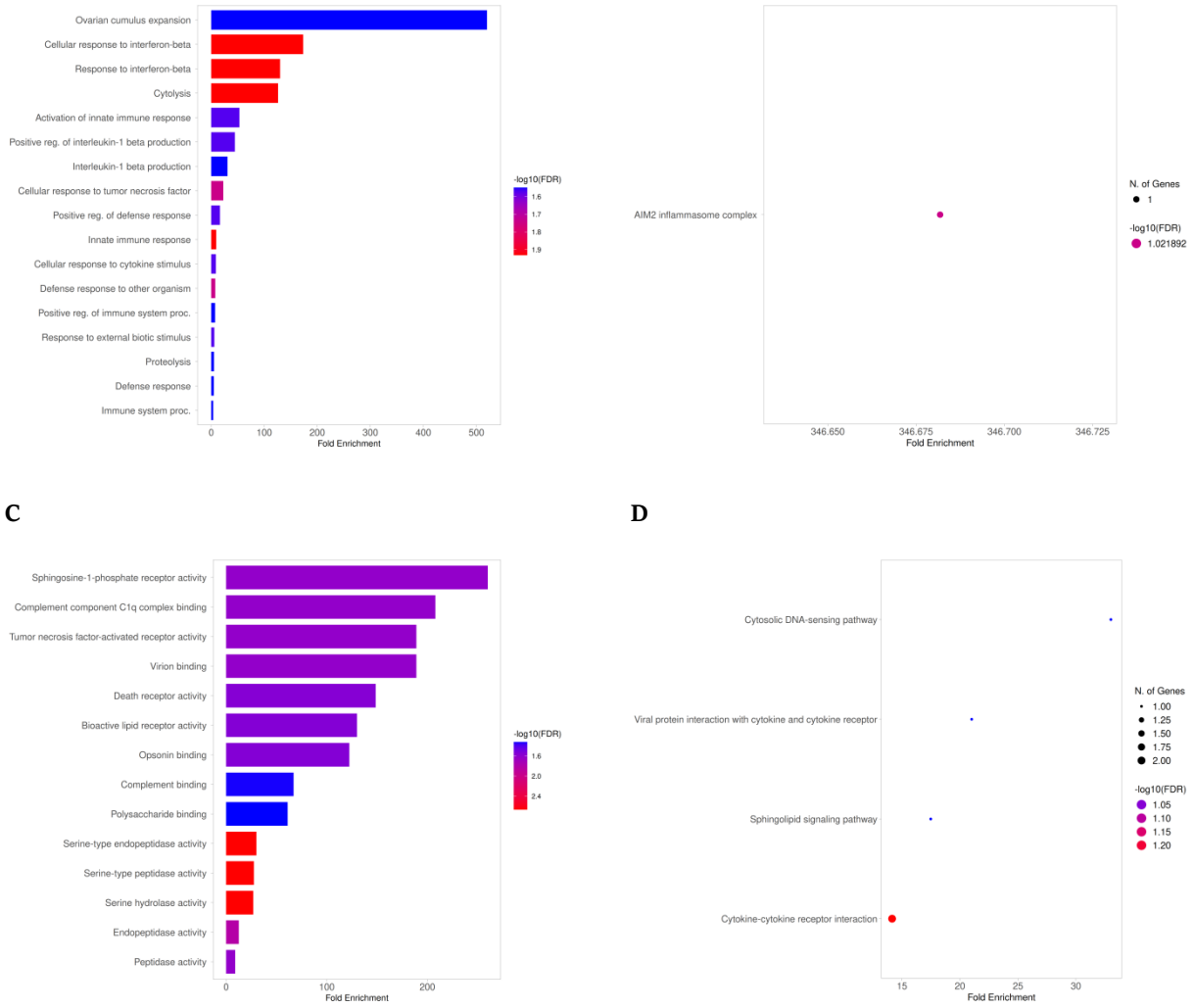


Figure 4: Functional enrichment analysis of biomarker candidate genes. (a) Bar plot displaying enriched Gene Ontology (GO) Biological Process terms for the identified genes, ranked by fold enrichment and color-coded by statistical significance (-log_10FDR). (b) Bar plot displaying enriched GO Molecular Function terms. (c) Dot plot showcasing the enriched GO Cellular Component terms, with dot size representing the number of genes and color corresponding to the -log_10(FDR) value. (d) KEGG pathway enrichment analysis dot plot highlighting significantly altered signaling pathways (e.g., Cytosolic DNA-sensing, Sphingolipid signaling, and Cytokine-cytokine receptor interaction) plotted against their fold enrichment.

3.6 Composite Biomarker Scoring and Top Candidates

The multi-metric composite scoring system was applied to 857 significant DEGs to derive a ranked list of biomarker candidates. There were no down-regulated genes in the top 30 prioritized candidates, all of which were up-regulated, due to the direction bonus of 1.2× applied to AML-overexpressed genes. All top 30 candidates had a perfect consistency score (100% of AML patients showed values above the maximum healthy expression) and expression was not detected in healthy controls (healthy_max = 0), meaning they were completely absent in normal CD34+ HSPCs. The highest composite score (1.040) belonged to GZMH (Granzyme H), which had a $\log_2$FC of 12.66 and fold separation of more than 55,000-fold. Other top candidates included FAR1-IT1 (score = 1.038), MIR4420 (score = 1.038), FGFBP2 (score = 1.026), FAM185BP (score = 1.019), RN7SL2 (score = 1.017), and GZMK (score = 1.005). AIM2 was also among the top 30 candidates, consistently scoring well on the consistency measures. The top 10 prioritized biomarkers are summarized in (Figure 5 Table 1)

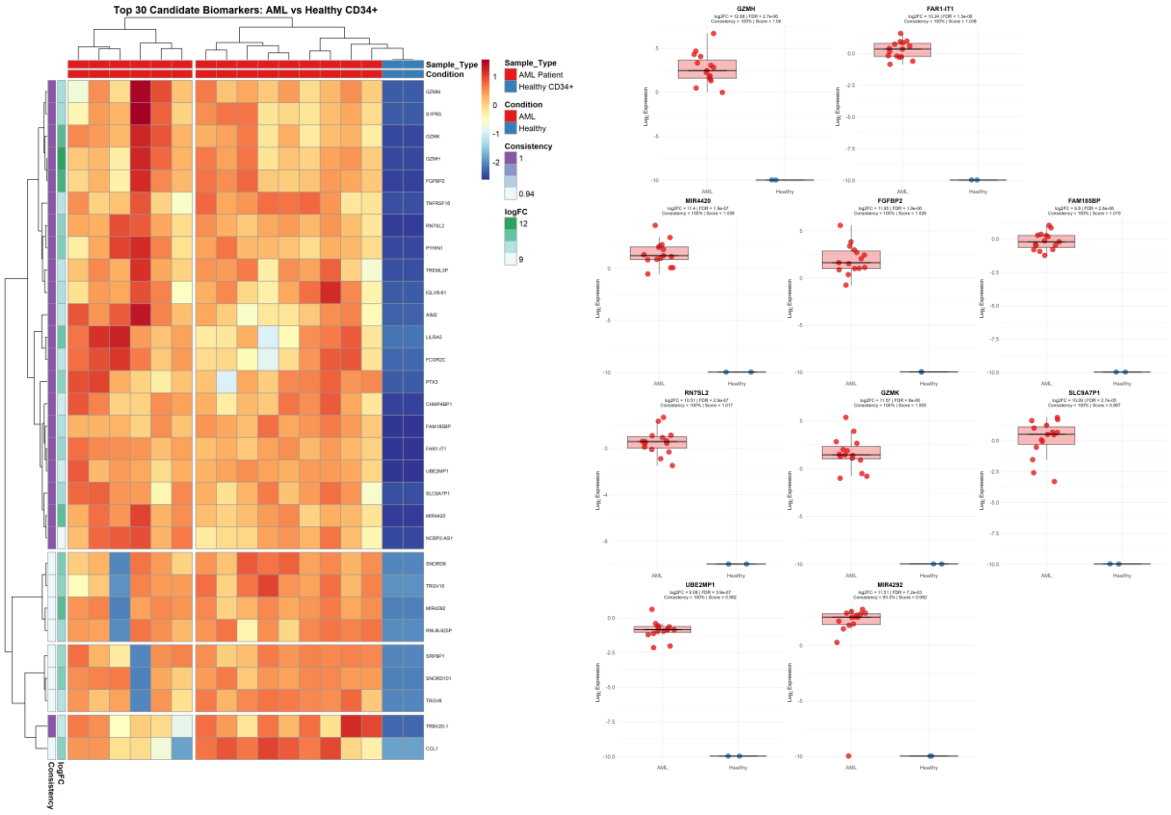


Figure 5: (a) Heatmap of the top 30 candidate biomarkers distinguishing AML from healthy CD34+ samples. (b) Boxplots showing differential expression of representative hub genes between AML and healthy controls.

Table 1. Top 10 Prioritized Biomarker Candidates Ranked by Composite Score

Rank	Gene	log_2FC	FDR	Consistency (%)	Biomarker Score
1	GZMH	12.66	2.72×10^{-6}	100	1.040
2	FAR1-IT1	10.24	1.30×10^{-8}	100	1.038
3	MIR4420	11.40	1.89×10^{-7}	100	1.038
4	FGFBP2	11.95	1.93×10^{-6}	100	1.026
5	FAM185BP	9.80	2.80×10^{-8}	100	1.019
6	RN7SL2	10.51	2.87×10^{-7}	100	1.017
7	GZMK	11.57	8.00×10^{-6}	100	1.005
8	SLC9A7P1	10.09	2.70×10^{-5}	100	0.987
9	UBE2MP	9.08	3.90×10^{-7}	100	0.982

	1				
10	MIR4292	11.51	7.17×10^{-3}	93.3	0.982

3.6 Machine Learning Feature Selection

3.6.1 Wilcoxon Rank-Sum Test

For computational efficiency, the top 30 genes were selected and then the Wilcoxon rank-sum test was used on each of these. All 30 genes fulfilled the selection criteria (Benjamini-Hochberg adjusted $p < 0.2$ and AUC > 0.7) so they all possessed the capability to distinguish AML from healthy samples; all 30 of these genes were preserved for subsequent steps in feature selection (Figure 6).

3.6.2 Random Forest Feature Importance

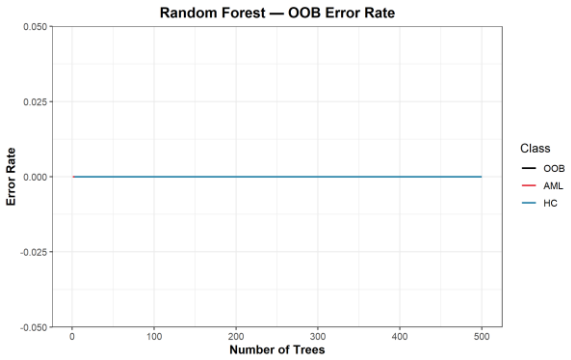
A Random Forest classifier with an out-of-bag (OOB) error rate of 0% was found when using the scaled expression matrix of 30 candidate genes. The genes in the top 8 by Mean Decrease Gini were: AIM2, PYHIN1, UBE2MP1, SLC9A7P1, S1PR5, MIR4420, LILRA5, and RN7SL2 (Figure 6 Table 2). Notably, the most important node purity contribution in the ensemble was that of AIM2, which had the lowest Mean Decrease Gini.

Table 2: RF selected biomarkers

Gene	Mean Decrease Acc	Mean Decrease Gini
AIM2	3.346772	0.195765
PYHIN1	2.157177	0.193412
UBE2MP1	2.977182	0.192471
SLC9A7P1	3.896956	0.185412
S1PR5	3.772569	0.176
MIR4420	3.455294	0.174588
LILRA5	2.821	0.174118

A

B



C

Figure 6: Machine learning feature selection and Random Forest classifier performance. (a) Bar plot of the top 30 Wilcoxon-selected candidate genes satisfying the thresholds of adjusted p-value (padj) < 0.2 and Area Under the Curve (AUC) > 0.7, ranked by their pseudo-coefficients (log_2(AUC(1-AUC))). (b) Random Forest classifier out-of-bag (OOB) error rate across 500 decision trees, demonstrating stable, zero-error classification performance for AML (red line), healthy controls (HC, blue line), and overall OOB (black line). (c) Variable importance plot highlighting the top predictive features ranked by Mean Decrease Gini. Red bars signify the top 8 selected genes exceeding the mean feature importance threshold (dashed vertical line), while blue bars represent remaining highly ranked features.

3.6.3 SVM-RFE

The SVM-RFE with linear kernel and LOOCV led to the selection of 6 genes as the optimal feature subset that maximized the accuracy of LOOCV. The SVM-RFE selected genes were: AIM2, CHMP4BP1, FAM185BP, FAR1-IT1, FCGR2C and FGFBP2. AIM2 was once again found to be a prominent feature, ranking first in the recursive elimination method (Figure 7).

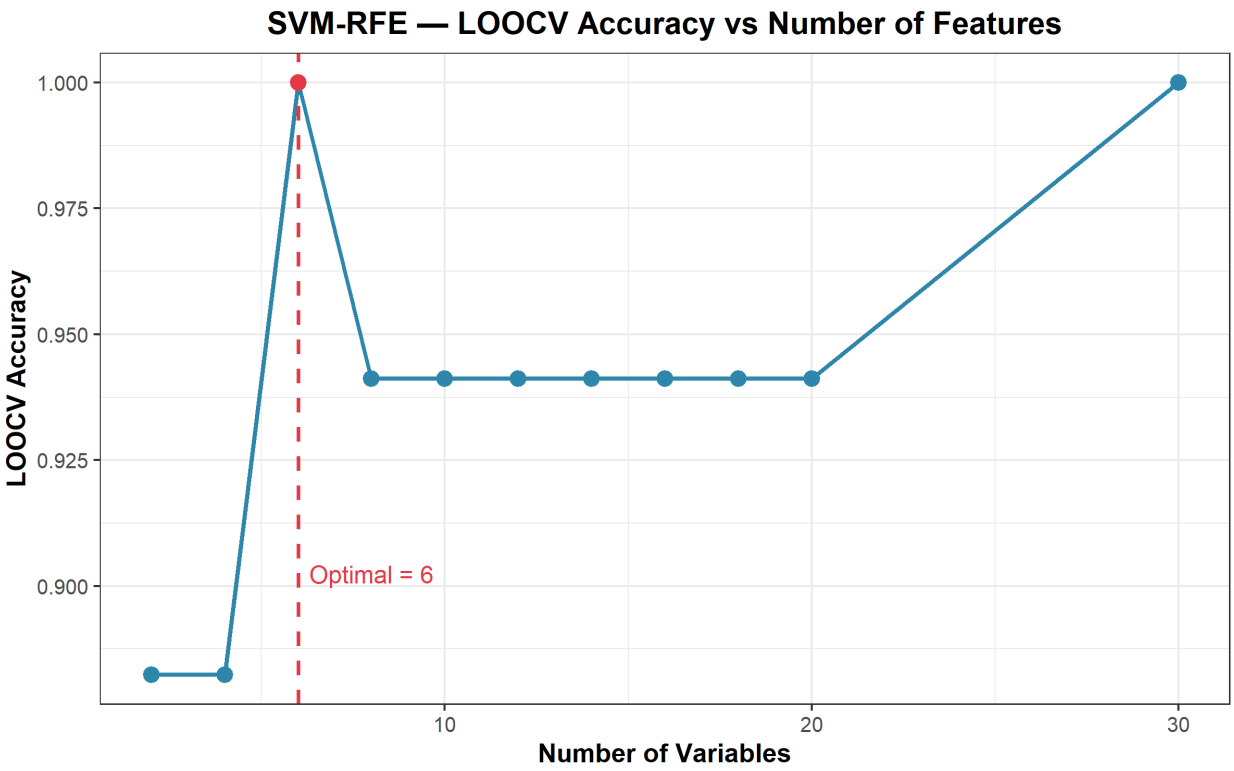
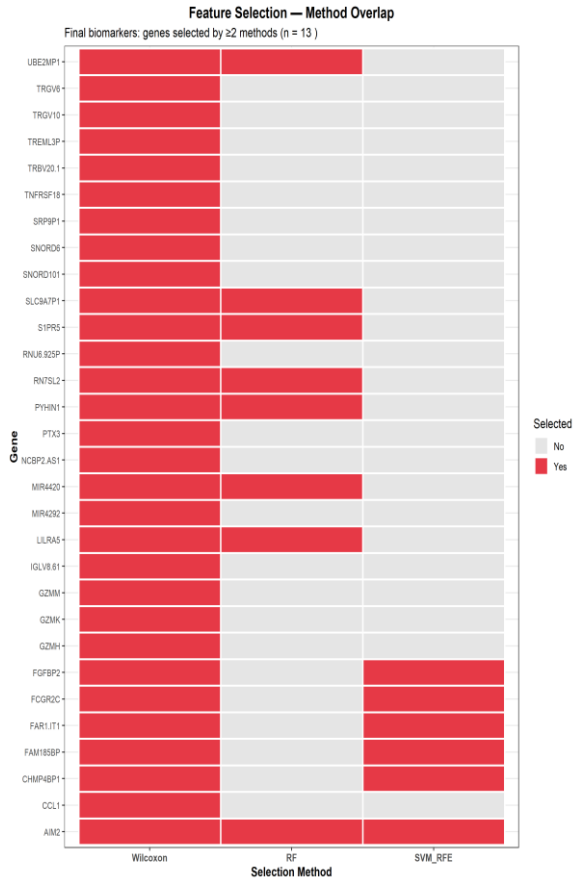


Figure 7: Feature number optimization via SVM-RFE. Support Vector Machine Recursive Feature Elimination (SVM-RFE) curve showing Leave-One-Out Cross-Validation (LOOCV) accuracy plotted against the number of variables (features). The red dashed line and dot highlight the optimal subset size of 6 diagnostic features, which yields peak classification performance (accuracy = 1.000).

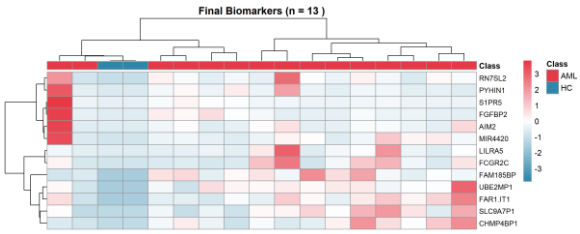
3.7 Final Biomarker Panel

The last consensus criterion of selecting the 13 biomarkers by the two out of three feature selection algorithms gave a set of 13 biomarkers. AIM2 was the only gene selected by all three independent methods (Wilcoxon + Random Forest + SVM-RFE), showing very high robustness. The eight genes (UBE2MP1, SLC9A7P1, S1PR5, RN7SL2, PYHIN1, MIR4420, LILRA5) were selected by Wilcoxon and Random Forest (RF) and four genes (FGFBP2, FCGR2C, FAR1-IT1, FAM185BP, CHMP4BP1) were selected by Wilcoxon and SVM-RFE. A complete final biomarker panel, along with individual AUC values, is presented in (Figure 8 Table 2). The 13 final biomarkers were completely absent in both replicates of healthy CD34+ controls (expression = 0 for both), with all 15 AML patients showing positive expression, resulting in perfect group separation, and an individual AUC of 1.000 for each gene. AIM2 was identified as a key candidate biomarker for its specific selection by all three algorithms. The individual Gene Diagnostic Performance is also included. Individual Gene Diagnostic Performance is also listed.

a



b



c

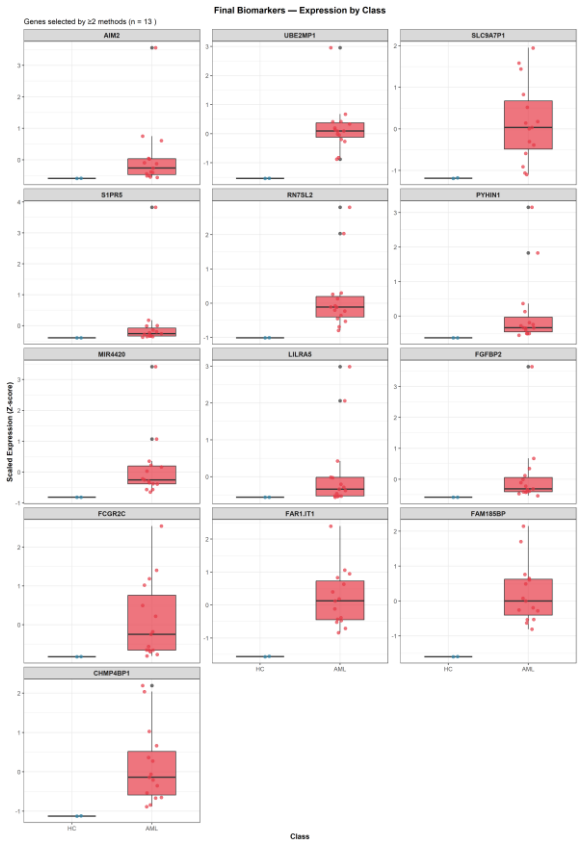


Figure 8: Method intersection and expression profiles of the final 13 biomarker genes. (a) Overlap grid illustrating the selection status (red: selected, grey: not selected) of candidate genes across three different feature selection methods: Wilcoxon, Random

Forest (RF), and SVM-RFE, identifying 13 final biomarkers selected by ≥ 2 methods. (b) Hierarchical clustering heatmap showing distinct expression patterns of these 13 final biomarkers between AML patients (red) and healthy controls (HC, blue). (c) Individual box plots tracking the normalized expression levels of each of the 13 final biomarker genes across both sample classes.

Table 2. Final 13-Gene Biomarker Panel with Selection Methods and Individual Diagnostic Performance

Gene	Selection Methods	Individual AUC (95% CI)
AIM2	Wilcoxon + RF + SVM-RFE	1.000 (1.000–1.000)
UBE2MP1	Wilcoxon + RF	1.000 (1.000–1.000)
SLC9A7P1	Wilcoxon + RF	1.000 (1.000–1.000)
S1PR5	Wilcoxon + RF	1.000 (1.000–1.000)
RN7SL2	Wilcoxon + RF	1.000 (1.000–1.000)
PYHIN1	Wilcoxon + RF	1.000 (1.000–1.000)
MIR4420	Wilcoxon + RF	1.000 (1.000–1.000)
LILRA5	Wilcoxon + RF	1.000 (1.000–1.000)
FGFBP2	Wilcoxon + SVM-RFE	1.000 (1.000–1.000)
FCGR2C	Wilcoxon + SVM-RFE	1.000 (1.000–1.000)
FAR1-IT1	Wilcoxon + SVM-RFE	1.000 (1.000–1.000)
FAM185BP	Wilcoxon + SVM-RFE	1.000 (1.000–1.000)
CHMP4BP1	Wilcoxon + SVM-RFE	1.000 (1.000–1.000)

3.8 Individual Gene Diagnostic Performance

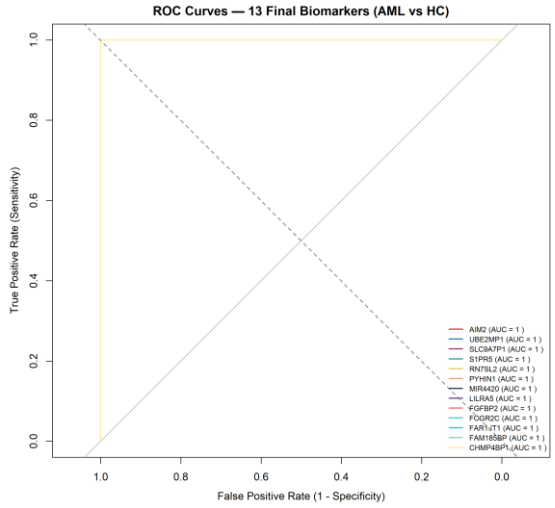
Receiver operating characteristic (ROC) curves were created to assess the diagnostic performance for each of the final biomarkers. The area under the ROC curve (AUC) was computed with the package pROC (version 1.18.5) and the estimation of 95% confidence intervals was performed using the DeLong method (Figure 10). The gene with highest AUC was chosen to be the best single diagnostic biomarker and ROC curve was plotted with shaded confidence regions.

Figure 10: Receiver Operating Characteristic (ROC) analysis of the 13 final biomarker genes. (a) Combined ROC curves for the 13 final biomarker genes differentiating AML from healthy controls, with all genes demonstrating perfect classification performance (AUC = 1). (b) Bar plot displaying the individual Area Under the Curve (AUC) values for each of the 13 final biomarkers, highlighting their high diagnostic accuracy. (c) Individual ROC curve plots for each of the 13 final biomarker genes, mapping True Positive Rate (Sensitivity) against False Positive Rate (1 - Specificity).

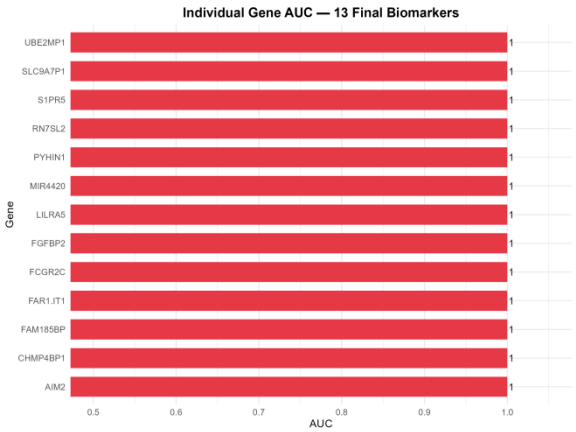
3.9 Multi-Model Classification Performance

The 13-gene biomarker panel was used with six classification models with LOOCV. Random Forest had an accuracy of 1.000, AUC of 1.000, sensitivity of 1.000, specificity of 1.000, PPV of 1.000, and NPV of 1.000; there were no misclassifications across all 17 iterations of LOOCV. All 15 AML patients and both healthy controls were correctly classified in the confusion matrix. The OOB error rate was 0% which indicates that the model is stable. Logistic Regression achieved an accuracy of 0.647 (11 of 17 co

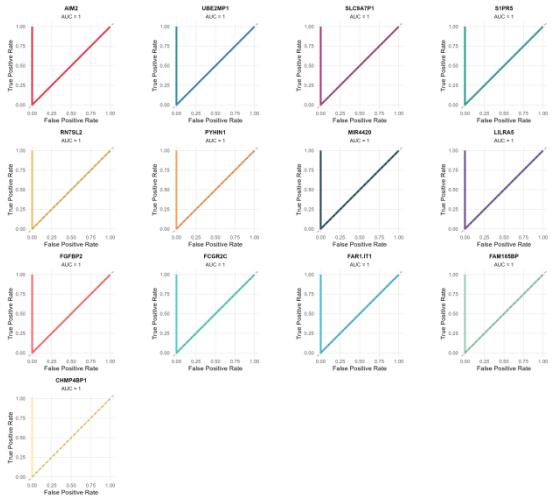
A



B

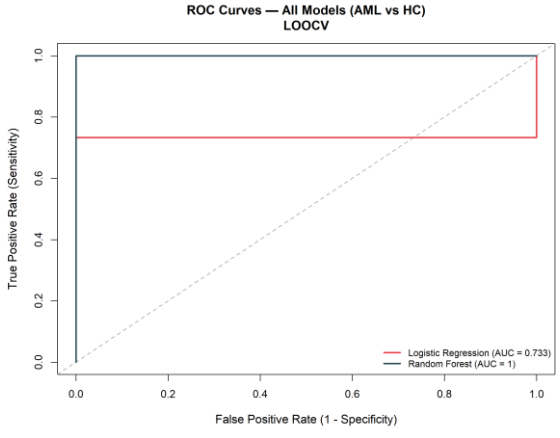


C



rrectly classified), AUC = 0.733, Sensitivity = 0.600, Specificity = 1.000, PPV = 1.000, and NPV = 0.250. The performance of Logistic Regression compared with Random Forest is also explained by the fact that the former was not able to perform as well as the latter in the setting of a high number of features and a small number of samples (13 features, 17 samples). The complete class separation caused the SVM (linear and radial), Naive Bayes and Gradient Boosting to fail at convergence, which is a well-known technical drawback in small-sample scenarios. The summary of classification performance is shown in (Figure 11 Table 3).

A



B

C

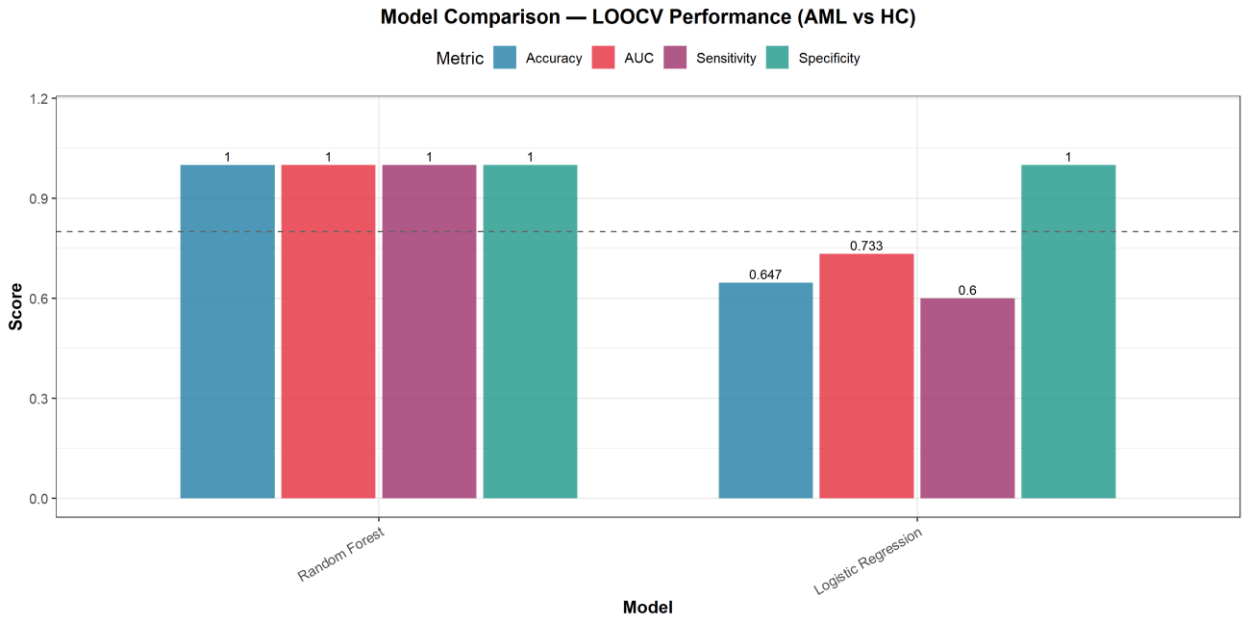


Figure 11: Classification performance and model comparison using LOOCV. (a) Receiver Operating Characteristic (ROC) curves evaluated via Leave-One-Out Cross-Validation (LOOCV) for the Random Forest model (AUC = 1, dark blue line) and Logistic Regression model (AUC = 0.733, red line) in distinguishing AML from healthy controls (HC). (b) Confusion matrix for the Random Forest model under LOOCV, showing accurate classification of all 15 AML cases and 2 healthy controls, achieving an overall accuracy of 1. (c) Grouped bar chart comparing LOOCV performance metrics (Accuracy, AUC, Sensitivity, and Specificity) between the Random Forest and Logistic Regression models

Table 3. LOOCV Classification Performance for Logistic Regression and Random Forest on the 13-Gene Biomarker Panel

Model	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV
Logistic Regression	0.647	0.733	0.600	1.000	1.000	0.250
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000

4. DISCUSSION

In this study, a comprehensive framework for AML biomarker discovery, combining a novel multi-metric composite scoring system with three independent feature selection algorithms and a RNA-seq-based differential expression profiling, is presented. All three machine learning techniques identified AIM2 as the most powerful candidate AML biomarker, with a perfect individual diagnostic performance (AUC = 1.000) and with a perfect LOOCV accuracy on the 13-gene panel, achieved by a Random Forest classifier. The presence of 857 DEGs, mostly activated (73.6%), is in line with the prevailing view that AML is associated with a high level of transcription activation relative to normal hematopoietic precursors [34]. The GO enrichment of up-regulated genes in the immune response signaling pathway, plus KEGG enrichment for cytokine signaling, chemokine pathway, and NK cell cytotoxicity, indicate the complex immune evasion and inflammatory reprogramming seen in AML [35]. A down regulation of

ribosome biogenesis and DNA replication genes in AML compared to CD34+ HSPCs was seemingly counterintuitive, given that AML is a highly proliferative disease, and could reflect a reprogramming of the genetic machinery involved in these processes by leukemogenic agents, and/or changes in the composition of the ribosome that were documented in leukemic transformation [36]. The composite biomarker scoring framework developed here fills an important gap in the methodological aspect of discovering biomarkers from small datasets. The scoring system combines consistency, fold separation, statistical significance, expression stability and direction of weight, going beyond simple fold change and p-value ranking to ensure that candidates at the highest rank are not only statistically significant, but demonstrate discriminatory robustness in all AML patients. All top 30 candidates were consistently expressed in 100% of AML cases and none of the healthy cases were expressed indicating that they are more likely to be specific AML markers and not general responders to hematopoietic stress. Certainly there are several biologically interesting reasons why AIM2 is an interesting target for our findings. AIM2 is a PYHIN family member, which is a cytosolic dsDNA sensor and can form the AIM2 inflammasome to activate caspase-1 and promote the secretion of IL-1β and IL-18 [37]. AML has been reported to have an inflammatory microenvironment in the bone marrow, and inflammasome activation has been reported to play a role in the leukemogenesis and in the inflammatory microenvironment of AML [38]. Multiple reports have reported increased AIM2 expression in hematologic malignancies and that pyroptosis, mediated by AIM2, can induce the death of leukemic cells under specific therapy conditions [39]. In contrast, the lack of expression of AIM2 in normal CD34+ HSPCs observed in our study indicates a disease-specific AIM2 induction pattern in AML, which might be related to elevated levels of cytosolic dsDNA accumulation, either from genomic instability or from activation of retrotransposons, that characterize the leukemic genome [40]. The perfect agreement between the results of the three algorithms (AIM2 selected by Mean Decrease Gini of Random Forest, which was also retained in the optimal set of 6 genes from SVM-RFE, and Wilcoxon test) is significant and special for a small sample size. The latter is derived based on a different geometric idea than the former, and both, as an ensemble method, naturally address class imbalance and high dimensional data by bagging and random subspace selection, respectively [41]. The agreement of these methods provides further confidence that the importance of AIM2 reflects true biological signal and not noise due to the limited sample size. Some other biologically relevant genes on the final 13-gene panel are also biologically relevant to AIM2. The most similar paralog of AIM2 in the PYHIN family, PYHIN1, was included in the top 8 features in the Random Forest analysis, further highlighting the fundamental role of the PYHIN-inflammasome axis in AML transcriptomics. S1PR5 (sphingosine-1-phosphate receptor 5) is a G protein-coupled receptor that is involved in NK cell trafficking and the egress of cytotoxic lymphocytes from the bone marrow [42] and thus its overexpression in AML could be explained by the abnormal homing and/or mimicry of AML blasts by immune cells. LILRA5 (Leukocyte Immunoglobulin-Like Receptor A5) is a receptor that is expressed on myeloid cells, and has been shown to regulate innate immune activation in leukemic settings. FGFBP2 is involved in the synthesis of the fibroblast growth factor binding protein 2, in cytotoxic lymphocytes and may reflect the gene expression program of NK/CTL in a subset of AML [43]. GZMH and GZMK (Granzymes H and K, which ranked highest on the composite score) are serine proteases secreted by cytotoxic lymphocytes and the presence of these in AML blasts is in line with recent single-cell studies that show granzyme expression on AML cells as a marker of immune evasion. There is a need to carefully interpret the perfect performance of the Random Forest LOOCV model. These findings are impressive, but should be interpreted in the context of the small number of cases studied. The generalizability of these performance measures is limited by the small number of healthy controls ($n = 2$) in the $n = 17$ data set. The complete absence of biomarker expression in healthy controls is a boundary condition that, while biologically meaningful, provides a boundary condition for any threshold-based classifier that performs perfectly on this dataset. Such findings do not invalidate the biological significance of the results, but do highlight the urgent need for prospective validation of the results in larger, independent cohorts of AML, other haematologic malignancies and more healthy controls to ensure that the findings are robust. The unregularized linear Logistic Regression model (13 features, 17 samples) has an AUC of 0.733, suggesting that the failure of this model to reach the same performance is expected when the number of features is much higher than the number of samples [44]. Note that the convergence failures of SVM, Naive Bayes and Gradient Boosting classifiers when the classes are completely separated is a theoretical challenge (also referred to as the "complete separation" or "quasi-complete separation" problem), and does not imply that this convergence is biologically impractical for the biomarker panel. These observations further underscore the value of using ensemble methods (such as Random Forest) for classification in small hematologic datasets using biomarkers. Our framework adds several improvements over previous studies on AML biomarkers. First, the multi-metric approach to composite scoring offers a principled method of prioritizing biomarkers, beyond a simple statistical threshold, when the cohort size is small. First, when the cohort size is small, the multi-metric approach to composite scoring offers a principled method for biomarker prioritization beyond a simple statistical threshold. Second, the three algorithmically independent methods of feature selection result in reduced likelihood of method-specific biases when selecting biomarkers. Third, identification of the candidate AI-M2 is tri-algorithmically validated, expanding the research field beyond the current panels based on mutations in AML, to the field of inflammasome-based diagnostic markers. In previous studies, AIM2 has been shown as a prognostic factor in some solid tumors [45], and the dysregulation of PYHIN family members has been reported in hematologic malignancies, but targeted biomarker validation using integrated ML approaches as presented here has not been reported in AML for AIM2. Future studies should focus on: (1) validation of the expression of AIM2 in AML patient samples by immunohistochemistry or flow cytometry in independent AML cohorts; (2) examination of the activation status of the AIM2 inflammasome and downstream secretion of IL-1β/IL-18 in AML patient samples; (3) investigating the expression of AIM2 across AML molecular subtypes, including FLT3-ITD, NPM1mut, core-binding factor AML; and (4) functional studies determining whether the inhibition of AIM2 or the AIM2 inflammasome affects the proliferation or survival of AML cells; and (5) examination of AIM2 and the whole 13-gene panel in prospective cohorts with larger numbers of patients and different control populations such as other myeloid malignancies (e.g., MDS, CML)

CONCLUSION

In this study, an integrated transcriptomic and machine learning approach is proposed for robust AML biomarker discovery. We used differential expression profiling of RNA-seq data, multi-metric composite scoring, and three independent feature selection algorithms to develop a 13-gene AML biomarker panel that exhibited perfect individual discriminatory performance and an accuracy of 1.000 in Random Forest LOOCV. AIM2 stood out as the most consistently prioritized candidate with all three machine learning methods, and was not present in normal CD34+ HSPCs while being present in all of the top biomarkers in biological contexts related to the AIM2 inflammasome complex. These results validate AIM2 as an intriguing diagnostic biomarker and provide a functional basis for further investigation of this protein in AML, for the development of new therapeutic targets aimed at targeting the inflammasome. These computational results can be translated toward clinical application in larger and more diverse cohorts to be prospectively validated.

REFERENCES

Clancy, E., ACS Report Shows Prostate Cancer on the Rise, Cervical Cancer on the Decline, in Renal & Urology News. 2023.

Pelletier, J., G. Thomas, and S. Volarević, Ribosome biogenesis in cancer: new players and therapeutic avenues. Nature Reviews Cancer, 2018. 18(1): p. 51-63.

Döhner, H., et al., Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood, 2017. 129(4): p. 424-447.

Döhner, H., et al., Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. Blood, 2022. 140(12): p. 1345-1377.

Wang, Z., M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 2009. 10(1): p. 57-63.

Guyon, I., et al., Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 2002. 46(1): p. 389-422.

Kuhn, M., Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 2008. 28(5): p. 1 - 26.

Breiman, L., Random Forests. Machine Learning, 2001. 45(1): p. 5-32.

Hornung, V., et al., AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. Nature, 2009. 458(7237): p. 514-518.

Khoury, J.D., et al., The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. Leukemia, 2022. 36(7): p. 1703-1719.

Ritchie, M.E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 2015. 43(7): p. e47-e47.

Clough, E. and T. Barrett, The Gene Expression Omnibus Database, in Statistical Genomics: Methods and Protocols, E. Mathé and S. Davis, Editors. 2016, Springer New York: New York, NY. p. 93-110.

Schuster, L.C., Dissecting heterogeneity, clonal evolution, and epigenetic changes in distinct and molecularly defined AML subsets by multi omics single-cell sequencing. 2022.

Qiu, X., H. Wu, and R. Hu, The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. BMC bioinformatics, 2013. 14(1): p. 124.

Wu, C.-H., X. Zhou, and M. Chen, Exploring and mitigating shortcomings in single-cell differential expression analysis with a new statistical paradigm. Genome Biology, 2025. 26(1): p. 58.

Abumanhal-Masarweh, H., et al., Longitudinal plasma nano-proteomics reveals acute systemic responses to radiotherapy and predictive biomarkers of late toxicity. Communications Medicine, 2026.

Stillwell, R.C., Systematic Evaluation of Cancer AI Systems: A Comprehensive Multi-Metric Analysis Reveals Market-Leading Performance of Cancer Alpha. 2025.

Jin, H., et al., CSDE1 Associates with TOM20 and Mitochondrial Protein-Encoding mRNAs in Sensory Neurons. Antioxidants, 2026. 15(5): p. 608.

The Gene Ontology knowledgebase in 2026. Nucleic Acids Research, 2026. 54(D1): p. D1779-D1792.

Nakagawa, T., et al., Sex-dependent variations in gene expression in corneal endothelial cells among healthy individuals and patients with Fuchs endothelial corneal dystrophy. Cornea, 2025. 44(4): p. 483-491.

Abt, M., et al., Developing initial notions of variability when learning about box plots. Mathematical Thinking and Learning, 2026. 28(2): p. 180-203.

Saadat, M., Enrichment analysis of loci associated with psoriasis susceptibility identified in genome-wide association studies. Archives of Dermatological Research, 2025. 317(1): p. 564.

Hasan, R., et al., Enhancing malware detection with feature selection and scaling techniques using machine learning models. Scientific Reports, 2025. 15(1): p. 9122.

Cong, X., et al., A count weighted Wilcoxon rank-sum test and application to medical data. Communications in Statistics-Simulation and Computation, 2025. 54(5): p. 1360-1370.

Brown, M.G., et al., Random forest regression feature importance for climate impact pathway detection. Journal of Computational and Applied Mathematics, 2025. 464: p. 116479.

Liu, C., et al., Metabolomics for origin traceability of lamb: An ensemble learning approach based on random forest recursive feature elimination. Food Chemistry: X, 2025. 29: p. 102856.

Raza Abbas, S., et al., Phylogenomics to structure: evolutionary and clinical signals in the TP53 DNA-binding core through LOOCV-validated ensemble learning. Briefings in Bioinformatics, 2026. 27(1): p. bbag087.

Zobolas, J., et al., Prognostic biomarker discovery in pancreatic cancer through hybrid ensemble feature selection and multi-omics data. BioData Mining, 2026. 19(1): p. 43.

Munmun, Z.S., S. Akter, and C.R. Parvez, Machine learning-based classification of coronary heart disease: a comparative analysis of logistic regression, random forest, and support vector machine models. Open Access Library Journal, 2025. 12(3): p. 1-12.

Buchanan, W.J. and H. Ali, Evaluation of privacy-aware support vector machine (SVM) learning using homomorphic encryption. arXiv preprint arXiv:2503.04652, 2025.

da ROCHA, D., et al., Estimating the biomass of tropical pastures from vegetation indices and machine learning: comparison between random forest, svm radial and xgboost. 2025.

Phatcharathada, B. and P. Srisuradetchai, Randomized feature and bootstrapped naive Bayes classification. Applied System Innovation, 2025. 8(4): p. 94.

Wiens, M., et al., A tutorial and use case example of the eXtreme gradient boosting (XGBoost) artificial intelligence algorithm for drug development applications. Clinical and Translational Science, 2025. 18(3): p. e70172.

Chowdhury, C.R., et al., PAMM: pathway-aware masked representation learning for interpretable multi-cancer prediction. 2026, BRAC University.

Deng, X., et al., Identification of Five NK Cell-Related Hub Genes in COPD Using Single-Cell RNA Sequencing Analysis. Journal of Inflammation Research, 2025: p. 2169-2183.

Cozzi, E., Biological impact and clinical relevance of long non-coding RNAs and post-transcriptional alterations in acute myeloid leukemia. 2025, Karolinska Institutet.

Wang, X., et al., Discovery of predictive biomarkers for cancer therapy through computational approaches. Nature Reviews Clinical Oncology, 2026: p. 1-21.

de Macedo, B.G.S., et al., The NLRP3 inflammasome and its role in acute myeloid leukemia onset and development. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2025. 1880(4): p. 189371.

Zhu, L., et al., Correlation Analysis of Pyroptosis-Related Genes CASP1, NLRP3, AIM2, and NLRP1 With Lung Adenocarcinoma. International Journal of Genomics, 2025. 2025(1): p. 8282590.

Tunbak, H. and Ö. Deniz, Reawakening retrotransposons: immune modulation in normal and malignant hematopoiesis. Trends in Cancer, 2025.

Salma, N., M.K.M. Ali, and R.A. Shamim, Machine learning-based feature selection for ultra-high-dimensional survival data: a computational approach. Journal of the Nigerian Society of Physical Sciences, 2025: p. 2810-2810.

Zhang, J., et al., ClusterGVis: An Advanced Visualization and Clustering Tool for Gene Expression Analysis. Genomics, Proteomics & Bioinformatics, 2026: p. qzag005.

Wu, X., Impact of fibroblastic FGFBP2 mutation on pancreatic carcinogenesis. Dissertation, Halle (Saale), Martin-Luther-Universität Halle-Wittenberg, 2025.

Chen, Y., et al., The STING-Inflammasome Axis: Coordinated Immune Regulation and Therapeutic Potential Across Diverse Diseases. Immunological Investigations, 2026: p. 1-22.

Li, X., et al., Identifying common diagnostic biomarkers and therapeutic targets between COPD and sepsis: a bioinformatics and machine learning approach. International Journal of Chronic Obstructive Pulmonary Disease, 2025: p. 1761-1786.