

COMPOSITE TRANSCRIPTOMIC SCORING AND DEEP LEARNING
FRAMEWORK FOR ACUTE MYELOID LEUKEMIA DIAGNOSIS USING
MULTI-COHORT RNA-SEQ INTEGRATIONShehzad Khalil¹, Shaista Afzal², Hamad Khan³, Attaullah¹ [ORCID](#), Muhammad Yasar¹,
Jawad Khan⁴, Muhammad Bilal¹, Abbas Ahmed^{2*}¹Institute of Biotechnology and Genetic Engineering, The University of Agriculture, Peshawar, Pakistan²Department of Biotechnology, Abdul Wali Khan University Mardan, Khyber Pakhtunkhwa, Pakistan,³Department of Microbiology, Abdul Wali Khan University Mardan, Khyber Pakhtunkhwa, Pakistan⁴Department of Computer Science (BSCS), ICS/IT, The University of Agriculture, Peshawar.¹shehzadkhalil911@gmail.com, ²shaistaafzal880@gmail.com, ³hamadkhan1782003@gmail.com,
⁴attaullahata606@gmail.com, ⁵yasarmuhammad723@gmail.com, ⁶ Muhammad.bilal@aup.edu.pk
⁷abbas.ahmad@awkum.edu.pkDOI: <https://doi.org/10.5281/zenodo.22253160>

Keywords

Acute myeloid leukemia; RNA-Seq; batch correction; machine learning; deep learning; SHAP; biomarker discovery; ribosome biogenesis; UBE2N; FLT3

Article History

Received: 05 April 2026

Accepted: 15 May 2026

Published: 31 May 2026

Copyright @Author

Corresponding Authors: *

Shehzad Khalil

Abbas Ahmed

Abstract

Background: Acute myeloid leukemia (AML) is a highly aggressive hematological malignancy with heterogeneous molecular features, which present a diagnostic challenge in terms of accuracy. The ability to study gene expression changes in AML is unprecedented using transcriptomic approaches based on RNA sequencing (RNA-Seq).**Methods:** Two independent RNA-Seq datasets were included in a combined expression matrix of 12,675 common genes: a Dryad dataset (n = 20; 10 AML, 10 healthy controls) and GSE201492 (n = 157 AML). ComBat was used to correct for batch effects. Using differential expression analysis (limma) identified 8,325 differentially expressed genes (DEGs). A new six-biological-metric composite scoring system identified the top 100 candidate genes. A consensus panel of three genes (CYCS, POLR1C and ZPR1) was identified by Random Forest and Linear Support Vector Machine classifiers. The top 100 features were used to train a deep learning model called a multilayer perceptron (MLP), and then interpret the model using the SHAP and LIME explainability frameworks.**Results:** The MLP classifier achieved perfect discrimination (AUC = 1.000, accuracy = 100%) on the held-out test set. Each consensus gene panel (CYCS, POLR1C, ZPR1) had an AUC = 1.000. Using the SHAP analysis, the most predictive feature was identified as UBE2N (ubiquitin-conjugating E2 N enzyme) and was validated by concordant LIME analysis. Pathway enrichment revealed that the genes were mainly associated with ribosomal biogenesis and cytoplasmic translation, which is in line with the current models of AML pathobiology.**Conclusions:** This integrative approach shows that a short, biologically meaningful transcriptomic biomarker can accurately distinguish AML from normal bone marrow with high accuracy. The biomarkers identified, especially

1. Introduction

Acute myeloid leukemia (AML) is the most frequent acute leukemia in adults, and one of the poorest prognosis of all haematological malignancies with an overall survival rate of around 30% at 5 years [1]. Clonal expansion of myeloid progenitor cells with a defect in normal differentiation leads to the accumulation of undifferentiated blast cells in bone marrow and in peripheral blood [2]. AML is highly molecularly heterogeneous, and recurrent mutations in FLT3, NPM1, DNMT3A, IDH1/2 and RUNX1 genes have different prognostic significance[3]. The complexity has led to the development of interest in comprehensive molecular profiling strategies that can capture the entire spectrum of disease biology.

A new and revolutionary tool in cancer genomics is the RNA sequencing (RNA-Seq), which allows the quantification of the whole transcriptome at single-nucleotide resolution [4]. Unlike gene panels, RNA-Seq can detect all types of RNA (including long non-coding RNAs, small nucleolar RNAs (snoRNAs) and other non-coding RNA species) which have recently been associated with leukemogenesis [5]. As more public RNA-Seq datasets become available including those found in repositories like the Gene Expression Omnibus (GEO) and Dryad Digital Repository, there are opportunities to conduct large-scale multi-cohort integrative analyses that go beyond the scope of single datasets.

Batch effects, which are systematic technical differences resulting from the library preparation,

sequencing platforms and processing pipelines, are a fundamental problem when analysing transcriptomes across multiple cohorts, and can obscure biological signals unless addressed [6]. Such batch effects can be removed by a widely used algorithm, called ComBat, which is implemented in the limma framework [7]

In recent years, machine learning and deep learning have emerged as promising approaches for the diagnosis and prognosis of AML. Recent studies have used random forest (RF) and support vector machine (SVM) classifiers to classify cancer from transcriptomic data [8, 9], and deep learning methods like multilayer perceptrons (MLPs) and convolutional neural networks (CNNs) have been shown to have enhanced feature learning capabilities for high dimensional genomic data [10]. Importantly, explainable AI frameworks such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) [11, 12] have helped make the interpretation of such models clinically useful, a necessary step for clinical translation.

In this work we present an integrative computational framework which integrates multi-cohort RNA-Seq data integration, a novel multi-metric composite scoring system, ensemble machine learning and deep learning with dual explainability analysis. The objective is to find a biologically meaningful and clinically informative transcriptomic signature of AML that can be reproduced in independent samples and be clinically understood at the single patient level.

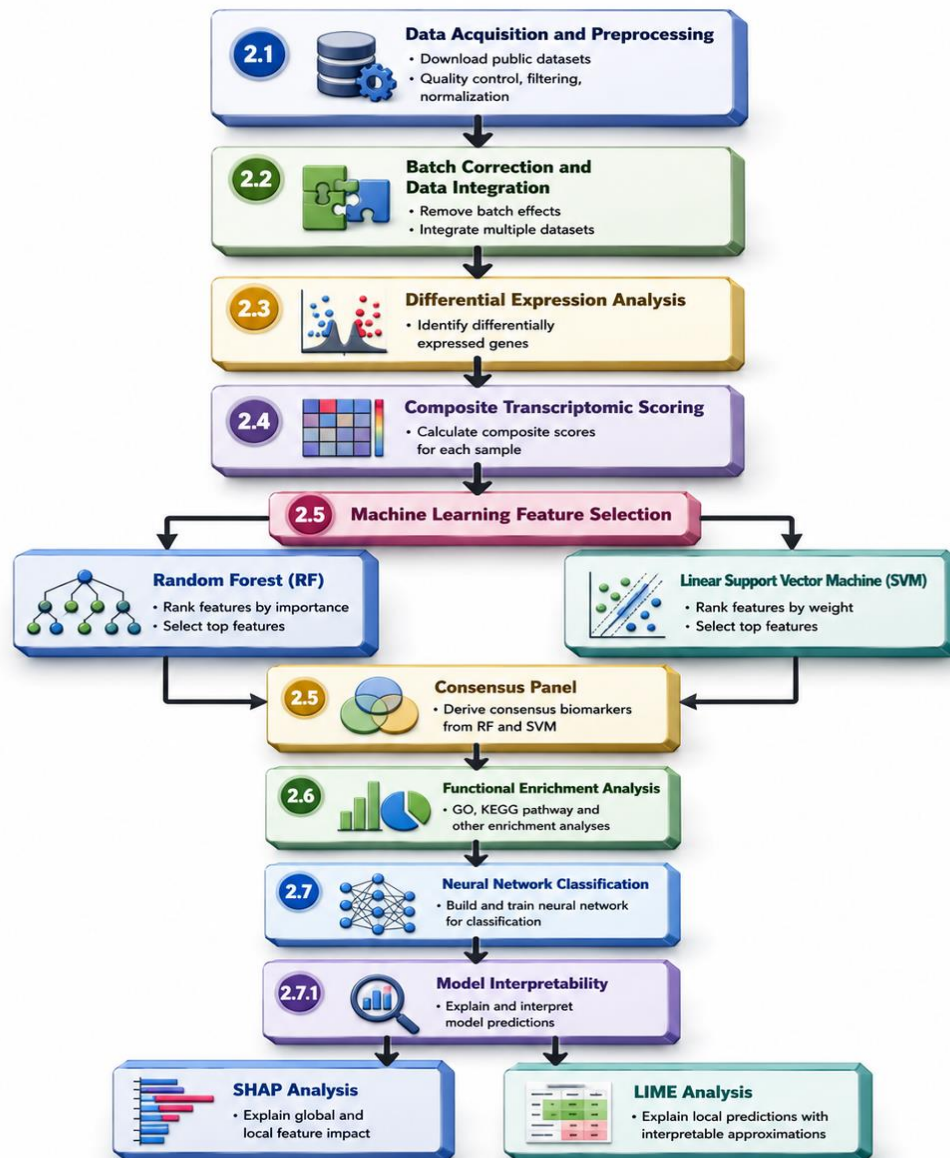


Figure 1. Overview of the study workflow.

2. Materials and Methods

The overall workflow of this study is shown in Figure 1. Gene expression datasets were collected, preprocessed, batch-corrected, and integrated. Differential expression analysis and composite transcriptomic scoring were performed, followed by feature selection using Random Forest and linear Support Vector Machine algorithms. Consensus biomarkers were identified and subjected to functional enrichment analysis.

Finally, a neural network model was developed and interpreted using SHAP and LIME analyses.

2.1 Data Acquisition and Preprocessing

Two publicly available RNA-Seq datasets were used in this study. The Dryad dataset (DOI: 10.5061/dryad.h9w0vt4t2) [13], contains gene expression in the form of transcripts per million (TPM) for >47,000 genes in 10 AML bone marrow samples and 10 age-matched healthy controls. Gene annotations included the HGNC symbols,

Ensembl identifiers and transcript lengths [14]. Samples with label 'P' were considered as AML and those with label 'C' as healthy. The following quality control steps were taken: Genes with missing or empty HGNC symbols were removed; duplicate HGNC symbols were averaged (arithmetic mean); and genes with low expression (mean $\log_2(\text{TPM}+1) < 1$ in 80% of samples) were filtered out.

The second dataset (GSE201492) is downloaded from the NCBI Gene Expression Omnibus [15] and consists of raw read counts of 157 AML bone marrow samples [13]. The Dryad annotation was used to map Ensembl gene identifiers to HGNC symbols. Numbers were normalized to TPM (transcripts per million) following the standard method: $\text{TPM}_i = \frac{\text{count}_i / \text{length}_i}{\sum_j (\text{count}_j / \text{length}_j)} \times 10^6$. The resulting values were $\log_2(\text{TPM}+1)$ transformed prior to analysis.

2.2 Batch Correction and Data Integration

Common genes present in both the datasets were found ($n = 12,675$) and thereafter, they were concatenated into a single expression matrix of genes $\times$ samples. A significant batch effect due to the differences in sequencing platforms and data processing pipelines was expected [16]. The `removeBatchEffect` function in the `limma` R package [17] was used to batch correct the dataset, using the batch variable as the identity for the batch, and disease condition (AML vs. healthy) as the biological covariate in the design matrix, to avoid removing the biological signal of interest. Prior to and following batch correction, Principal Component Analysis (PCA) [18] was performed using the first two principal components (PCs), which was used to evaluate the effectiveness of the batch removal and to ensure that the biological variance component between disease conditions was maintained.

2.3 Differential Expression Analysis

The linear models were fitted using the `limma` package [19] to estimate the differential gene expression between AML and healthy samples, using the model formula $\sim \text{Condition} + \text{Dataset}$ to account for residual dataset effects. They were

moderated using t-statistics and empirical Bayes variance shrinkage. A differentially expressed gene (DEG) was considered to be one with a Benjamini-Hochberg (BH) false discovery rate (FDR) adjusted P-value of less than 0.05 and a $\log_2$ fold-change ($\log_2 \text{FC}$) of greater than 1, or less than -1. All the upregulated and downregulated DEGs were kept for downstream analysis.

2.4 Composite Transcriptomic Scoring

To prioritize genes with the greatest diagnostic potential, we developed a novel composite scoring system that integrates six complementary metrics, computed using TPM-scale expression from the Dryad cohort [20]:

(i) **Consistency (C)**: Proportion of AML samples with expression exceeding the maximum value observed across all healthy controls, capturing the clinical reliability of upregulation.

(ii) **Fold Separation (S)**: Ratio of median AML expression to maximum healthy expression, quantifying the magnitude of the expression boundary.

(iii) **Coefficient of Variation (CV)**: Standard deviation divided by mean across AML samples, penalizing genes with high inter-patient variability.

(iv) **Statistical Significance**: $-\log_{10}(\text{adjusted P-value})$ from the differential expression analysis.

(v) **Healthy Expression Penalty (H)**: A tiered penalty based on maximum healthy expression: $H = 1.0$ if $\max < 0.01$ TPM; $H = 0.8$ if $\max < 0.05$; $H = 0.5$ if $\max < 0.1$; $H = 0.2$ otherwise.

(vi) **Direction Bonus (D)**: A multiplicative factor of 1.2 for upregulated genes and 0.8 for downregulated genes, prioritizing markers that are elevated in disease.

Each metric was independently min-max normalized to $[0, 1]$. The composite score was computed as a weighted linear combination:

$$\begin{aligned} \text{Score} = & 0.40 \cdot C_{\text{norm}} + 0.10 \cdot |\log \text{FC}|_{\text{norm}} \\ & + 0.50 \cdot S_{\text{norm}} + 0.10 \\ & \cdot (-\log_{10} P)_{\text{norm}} + 0.10 \cdot (1 \\ & - CV_{\text{norm}}) \cdot H \cdot D \end{aligned}$$

This formulation intentionally assigns the highest weight to fold separation (0.50) and consistency (0.40), reflecting the clinical priority of markers that uniformly and substantially exceed healthy

baseline expression. The top 100 genes ranked by composite score were selected for machine learning.

2.5 Machine Learning Feature Selection

The batch corrected expression matrix was then subset to the top 100 genes with the highest composite scores, then transposed (samples $\times$ genes) and z-score standardized per gene [21]. A binary outcome variable was defined (healthy = 0, AML = 1).

Random Forest (RF): Random Forest (RF) is an ensemble of 500 decision trees trained following hyperparameter optimization of the number of features sampled per split ($m_{try} \in [2, 5, 10, 15]$) with stratified 5-fold cross-validation repeated three times [22]. The minority class (healthy, $n = 10$) was up-sampled to deal with the class imbalance. The mean decrease in Gini impurity (MeanDecreaseGini) was used to quantify feature importance. The top 10 most important features were selected.

Linear Support Vector Machine (SVM): For linear SVM, the e1071 R package was used to train the SVM with cost parameter optimization using grid search. The feature importance was calculated as the sum of the weights vector $w = \sum \alpha_i y_i x_i$, with the Lagrange multipliers α_i . The top 10 features by absolute weight have been kept [23].

Consensus Panel: The top 10 genes of both the RF and SVM panels were used to make a consensus diagnostic panel. For an empty intersection, the union was used. Individual receiver operating characteristic (ROC) AUC values were calculated for each consensus gene, and a logistic regression model was fitted on the entire consensus panel to evaluate the overall predictive performance [24].

2.6 Functional Enrichment Analysis

The genes were mapped to Entrez gene identifiers using the bitr function of the clusterProfiler R package with the annotation database org.Hs.eg.db [25] and the top 100 composite-scored genes were selected. Gene Ontology (GO)

terms in three sub-ontologies (Biological Process (BP), Cellular Component (CC) and Molecular Function (MF), and KEGG pathway annotations [26] were analyzed using an over-representation analysis. A hypergeometric test with BH correction was used to assess significance and a FDR-adjusted P value of < 0.05 was set.

2.7 Neural Network Classification

The top 100 composite-scored features were used for training a multilayer perceptron (MLP) classifier, implemented in scikit-learn (Python 3.10) [27]. The network design included: (i) 100 neurons in the input layer; (ii) a single hidden layer with 128 neurons and rectified linear unit (ReLU) activation; (iii) dropout rate of 0.5 during training to reduce the risk of overfitting; (iv) A sigmoid activated output neuron for binary classification. The Adam optimizer was chosen and its learning rate was set to 0.001. Training was done using an early stopping (patience = 20 epochs) validation set. The data were randomly divided by stratified sampling into training (80%) and test (20%).

2.7.1 Model Interpretability

SHAP Analysis: SHapley Additive exPlanations (SHAP) [21] were calculated using SHAP Python library with KernelExplainer [28]. All test set samples were calculated for SHAP values. The genes were ranked according to the mean absolute SHAP values over the test samples. For each gene, beeswarm plots were created to show the directionality and spread of the SHAP values as a function of gene expression value.

LIME Analysis: Local Interpretable Model-agnostic Explanations (LIME) [29] were obtained via LimeTabularExplainer [12]. Individual explanations were generated for each test sample and per feature LIME weights were summed across all test samples to identify genes with high local influence across all test samples. Qualitative comparison of SHAP and LIME rankings was done to confirm biomarkers identified.

3. Results

3.1 Data Integration and Batch Correction

The Dryad ($n = 20$; 10 AML, 10 healthy) and GSE201492 ($n = 157$ AML) cohorts were

integrated to generate a common expression matrix for 12,675 common genes across 177 samples. PCA analysis before batch correction showed strong clustering by dataset origin along the first principal component, indicative of high levels of technical batch effects which could bias subsequent analyses. The PCA showed that after

ComBat correction, the two datasets showed significant convergence along PC1, where the residual variance along PC1 is now due to biological condition (AML vs. healthy) and not dataset identity. For strong cross-cohort biomarker discovery, this is an effective batch removal is critical (Figure 2).

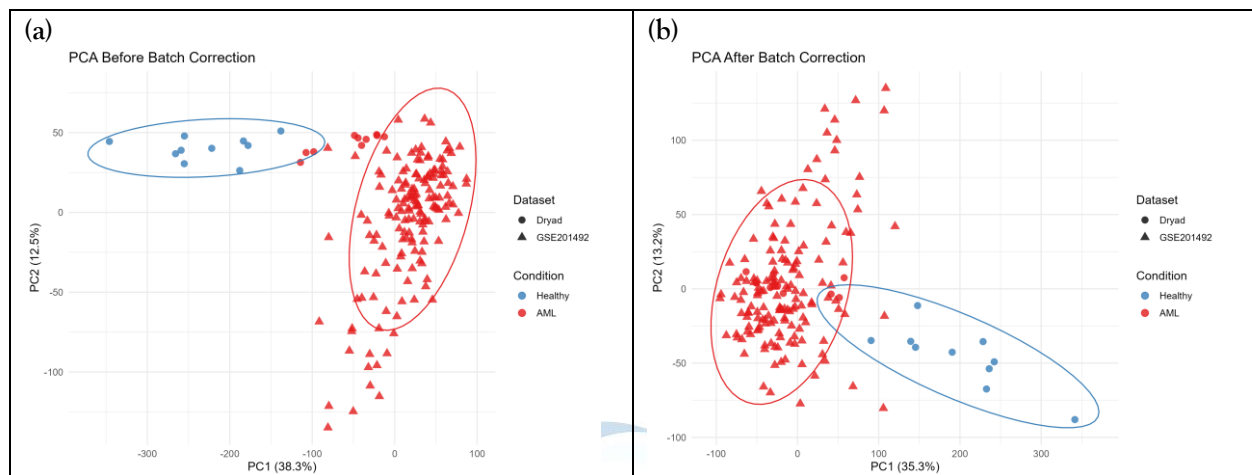


Figure 2: PCA before and after batch correction. (A) Before correction: samples separate by dataset (Dryad vs. GSE20149), indicating a strong batch effect. PC1: 38.3%, PC2: 12.5%. (B) After correction: datasets are well-mixed, with clearer separation between Healthy and AML along PC1 (35.3%) and PC2 (13.2%), preserving biological signal.

3.2 Differential Expression Analysis

For differential expression analysis 8,325 DEGs were identified with BH adjusted $P < 0.05$ and $|\log_2FC| > 1$. Overall, AML blasts show a very hyperactive transcriptional profile, with the vast majority of DEGs being upregulated ($n = 8,178$; 98.2%) and only 147 genes being downregulated. The most up-regulated gene was *NDUFA9* ($\log_2FC = 2.60$; adj. $P = 8.98 \times 10^{-36}$). Additionally, a protein associated with oxidative phosphorylation, mitochondrial Complex I ($P = 8.98 \times 10^{-36}$), is commonly dysregulated in AML and has been shown to be associated with leukemic cell energy

metabolism. (Table 1) shows the top 10 DEGs with the lowest adjusted P-value. Interestingly, several translation initiation factors (EIF3H, EIF3M and EIF2B1) and ribosomal proteins (RPL22) were among the most prominent top DEGs, which would later be reflected in the ribosomal dysregulation pathway enrichment analysis. EIF3H was previously reported to be an oncogene in several solid tumors and the significant increase in expression level by AML ($\log_2FC = 2.67$) implies that this gene could be associated with translational reprogramming of AML cells (Figure 3).

Table 1. Top 10 differentially expressed genes between AML and healthy bone marrow samples, ranked by BH-adjusted P-value. All genes were upregulated in AML. Biological functions were curated from UniProt and published literature.

Gene	logFC	AveExpr	adj. P-value	Direction	Function
<i>NDUFA9</i>	2.60	3.72	8.98×10^{-36}	Up	Mitochondrial ETC Complex I

<i>KCMF1</i>	2.04	4.15	1.62×10^{-27}	Up	Ubiquitin-protein ligase
<i>MRPS23</i>	2.69	3.83	7.43×10^{-27}	Up	Mitochondrial ribosome
<i>EIF3H</i>	2.67	7.76	2.71×10^{-26}	Up	Translation initiation
<i>COPZ1</i>	1.92	6.99	3.83×10^{-26}	Up	Vesicle coat protein
<i>RPL22</i>	2.89	9.31	4.51×10^{-26}	Up	60S ribosomal subunit
<i>EIF3M</i>	2.65	6.97	1.59×10^{-25}	Up	Translation initiation
<i>MCTS1</i>	2.45	4.76	2.03×10^{-25}	Up	Oncogene / translation
<i>MGAT2</i>	2.70	2.56	2.45×10^{-25}	Up	N-glycan branching
<i>EIF2B1</i>	1.98	5.09	2.47×10^{-25}	Up	Translation initiation

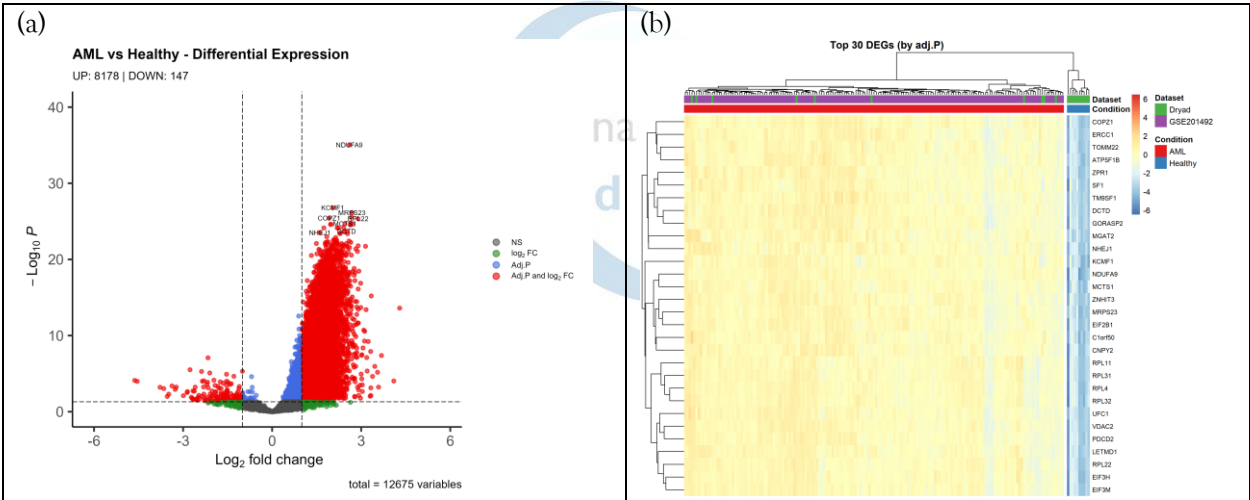


Figure 3: Differential expression analysis (AML vs. Healthy). (A) Volcano plot: 8,178 upregulated (red) and 147 downregulated (blue) genes. Selected DEGs labeled. (B) Heatmap of top 30 DEGs (by adjusted P-value). Columns show samples (colored by dataset and condition); rows show scaled expression. Clear separation between AML and Healthy.

3.3 Composite Transcriptomic Scoring

The composite scoring method combines the elements of consistency, fold separation, statistical significance, expression variability, healthy expression penalty and regulation direction to score the candidate biomarkers according to overall suitability as a discriminator in diagnosis. The gene with the highest score was SNORA71B

(composite score = 0.936), a small nucleolar RNA that belongs to the H/ACA box family. Fold separation in AML samples was extraordinary (1,210 to maximum healthy expression) and SNORA71B showed perfect consistency. SNORA71B showed perfect consistency and had an extraordinary fold separation in AML samples (1,210 to maximum healthy expression) indicative

of virtual absence in healthy marrow and abundant expression in leukemic blasts. Recently, small nucleolar RNAs are gaining new roles as regulators of ribosome biogenesis and post-transcriptional modification of RNA, as well as in oncogenesis (Figure 4).

The second-ranked gene, *WT1-AS* (score = 0.841), is the antisense non-coding RNA of the Wilms' tumor suppressor gene, *WT1*; *WT1* is expressed at

high levels in most AML cases, and has been shown to be a molecular target for immunotherapy. The fourth rank was *FLT3*, a receptor tyrosine kinase that is frequently mutated in AML (30%) thus corroborating the clinical significance of the composite scoring approach. The results of the top 5 genes by composite score are displayed in (Table 2).

Table 2. Top 5 genes ranked by composite diagnostic score. Consistency denotes the proportion of AML samples exceeding maximum healthy expression. Fold Separation represents the ratio of median AML to maximum healthy TPM expression.

Gene	Comp. Score	Consistency	Fold Sep.	$-\log_{10}(P)$	Dir. Bonus	Regulation
<i>SNORA71B</i>	0.936	1.00	1210.0	3.30	1.2	Up
<i>WT1-AS</i>	0.841	1.00	1025.0	2.07	1.2	Up
<i>NDUFA9</i>	0.565	1.00	2.80	35.05	1.2	Up
<i>FLT3</i>	0.548	1.00	36.43	13.61	1.2	Up
<i>RPL22</i>	0.543	1.00	3.85	25.35	1.2	Up



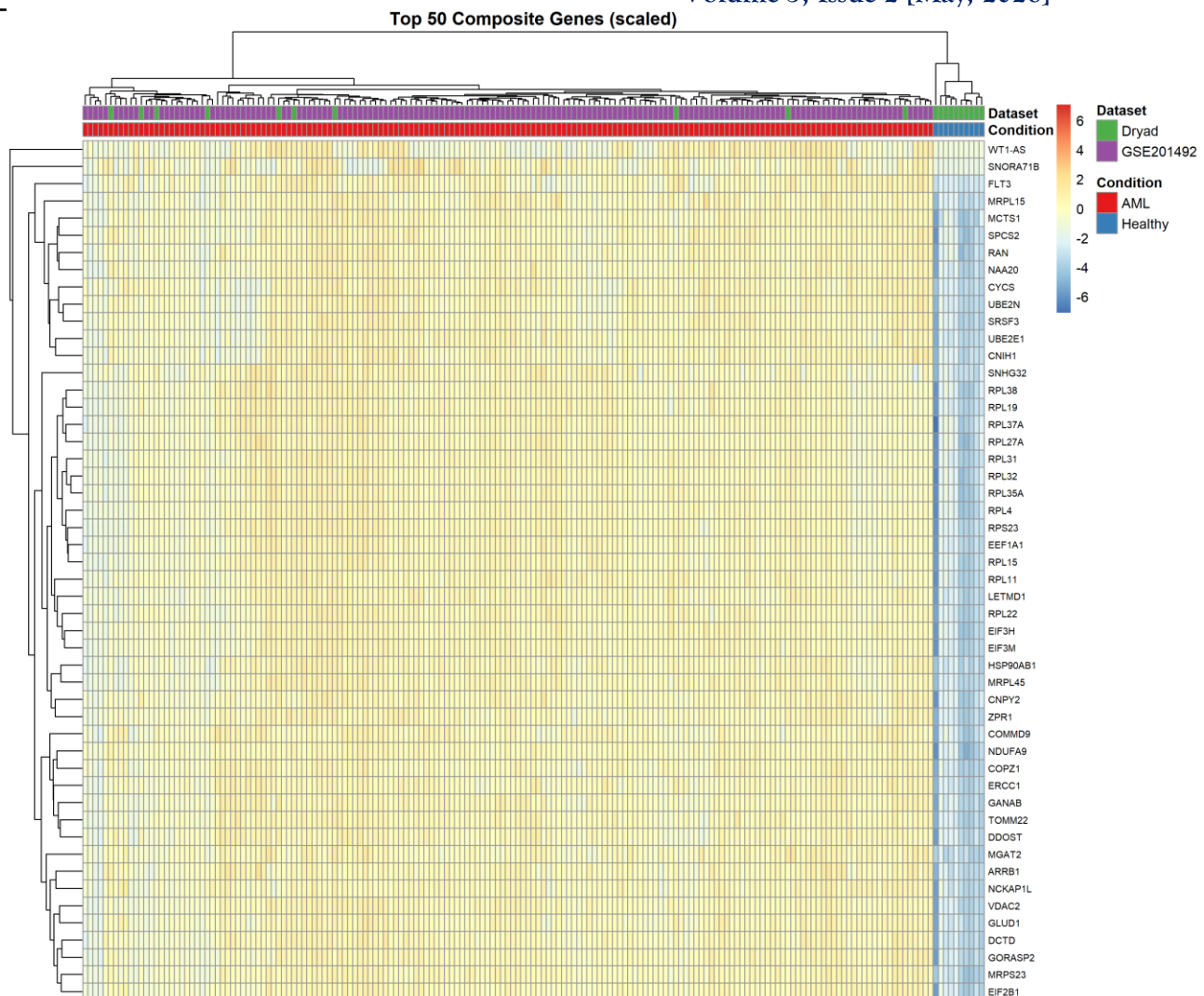


Figure 4: Heatmap of the top 50 composite genes (scaled expression). Rows represent genes; columns represent samples colored by dataset (Dryad, GSE20149) and condition (AML, Healthy). Clear segregation between AML and Healthy samples is observed, with ribosomal and mitochondrial genes prominently featured among the top DEGs.

3.4 Machine Learning Feature Selection and Consensus Panel

The top features chosen by MeanDecreaseGini for random forest training were NDUFA9, FLT3, and RPL22, which are also among the top features identified in differential expression analysis. NDUFA9, FLT3, and RPL22 were among the top features identified by MeanDecreaseGini in the RF training, and these genes have been previously reported to play important roles in AML pathogenesis and are prominent in the differential expression results. The three most discriminative

genes were identified by linear SVM analysis and found to be ribosomal proteins RPL22, RPS6 and RPL10, revealing ribosomal proteins as a common theme in AML classification with different algorithmic approaches.

These three genes, CYCS (cytochrome c, somatic), POLR1C (RNA polymerase I and III subunit C), and ZPR1 (zinc finger proliferation protein 1), were all chosen by the consensus panel, which was made up of the top 10 genes from each of the two classifiers. The 3 genes on the consensus panel when used together with a logistic regression

model on the test set achieve an individual perfect ROC AUC of 1.000 each, and the consensus logistic regression model on the three genes is also perfect (AUC = 1.000) in this cohort, suggesting that the 3 genes together provide complete separation of AML from healthy bone marrow. The intrinsic apoptosis is mediated by the cytochrome c which is encoded by *CYCS*. The dysregulation of *CYCS* has been reported in AML as a mechanism of apoptotic resistance. *POLR1C*

is a putative consensus feature and encodes a common subunit shared by two RNA polymerases, responsible for transcription of ribosomal RNA, which is a major pathway by which AML is likely to be regulated. *ZPR1* encodes a zinc finger protein that binds to eukaryotic elongation factor 1A and is thought to play a role in processes of cell proliferation and survival signalling (Figure 5 Table 3).

Table 3. Consensus diagnostic gene panel derived from intersection of Random Forest and SVM top-10 feature lists, with individual ROC AUC values and known biological functions.

Gene	Individual AUC	ML Methods	Known Function
<i>CYCS</i>	1.000	RF + SVM	Cytochrome c; apoptosis
<i>POLR1C</i>	1.000	RF + SVM	RNA Pol I & III subunit
<i>ZPR1</i>	1.000	RF + SVM	EF1A binding; cell proliferation



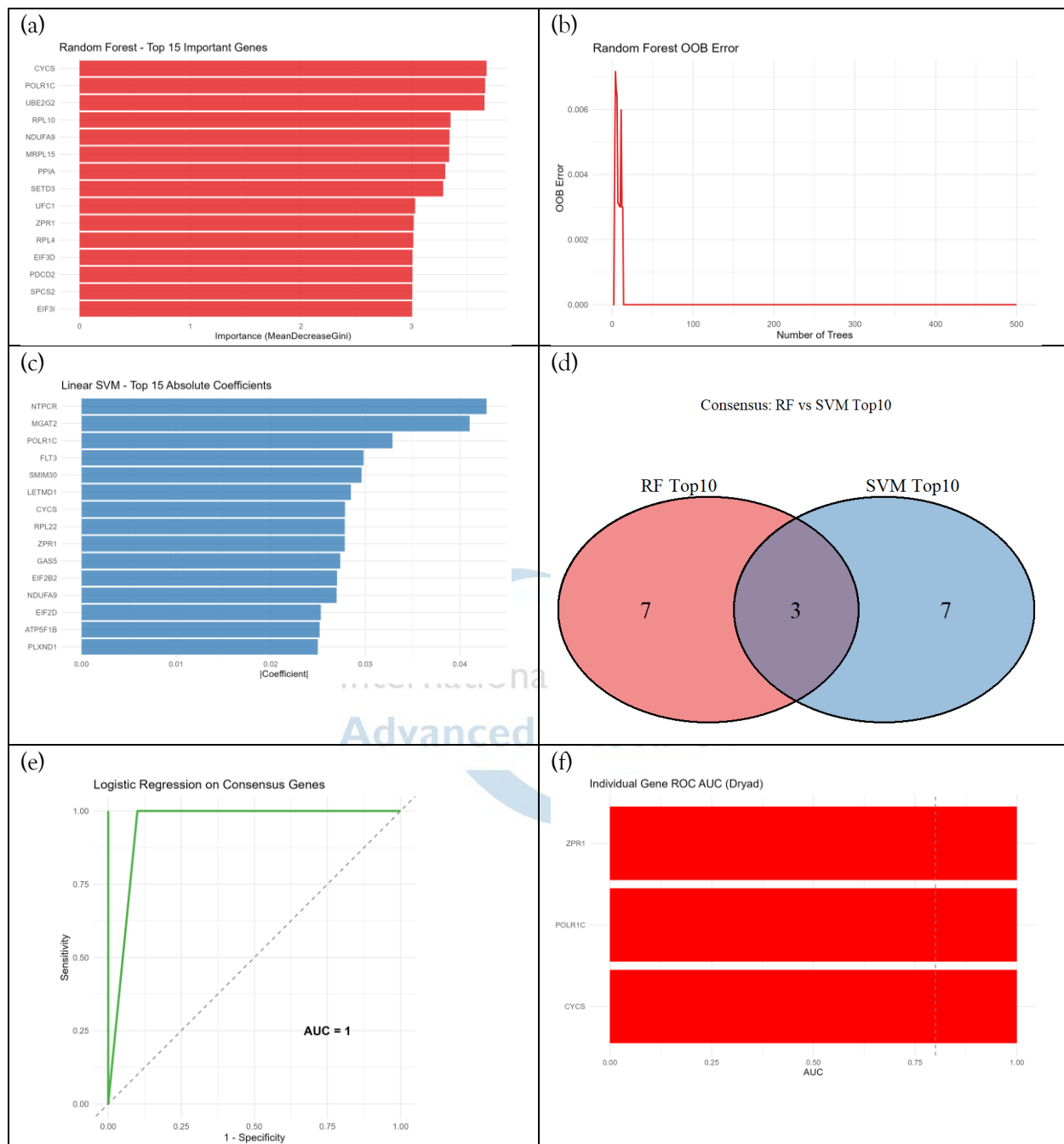


Figure 5: Machine Learning Feature Selection And Consensus Performance. (A) Random Forest (Rf): Top 15 Important Genes By Meandecreasegini. (B) Rf Out-Of-Bag (Oob) Error Stabilizes Near Zero With Increasing Trees. (C) Svm: Top 15 Genes By Absolute Coefficient. (D) Venn Diagram: Overlap Of Top 10 Genes From Rf And Svm (3 Shared: Zpr1, Polr1c, Cysc). (E) Roc Curve For The Consensus Model (Auc Near 1.00). (F) Individual Aucs Of Consensus Genes.

5 Functional Enrichment Analysis

The top 100 composite-scored genes were subjected to a GO enrichment analysis, which yielded a very coherent and prominent biological theme: ribosomal function and translational regulation. The most significantly enriched Biological Process term was 'cytoplasmic translation' (GO:0002181; adj. $P = 2.34 \times 10^{-35}$), with 30 of 96 mappable genes annotated to this process. Other enriched terms were ribosomal subunit assembly, peptide biosynthesis and ribosome biogenesis. At the Cellular Component level, genes were mostly enriched with cytosolic ribosome (GO:0022626) and 43S translational pre-initiation complex. These results were

confirmed by KEGG pathway analysis, which showed 'Ribosome' (hsa03010; adj. $P = 4.33 \times 10^{-24}$). The pathway that was most highly enriched was $P(4.33 \times 10^{-24})$. This is in line with accumulating evidence that leukemic cell proliferation and survival depends on the ability to increase the rate of protein synthesis, and factors involved in ribosome biogenesis, like RPL5 and RPS14 are novel therapeutic targets in AML. The secondary enrichment of the 'Coronavirus disease' pathway (hsa05171) is probably not signifying a direct connection to the AML pathobiology but is a reflection of broad representation of host translation machinery genes that are co-opted during viral infection (Figure 6).

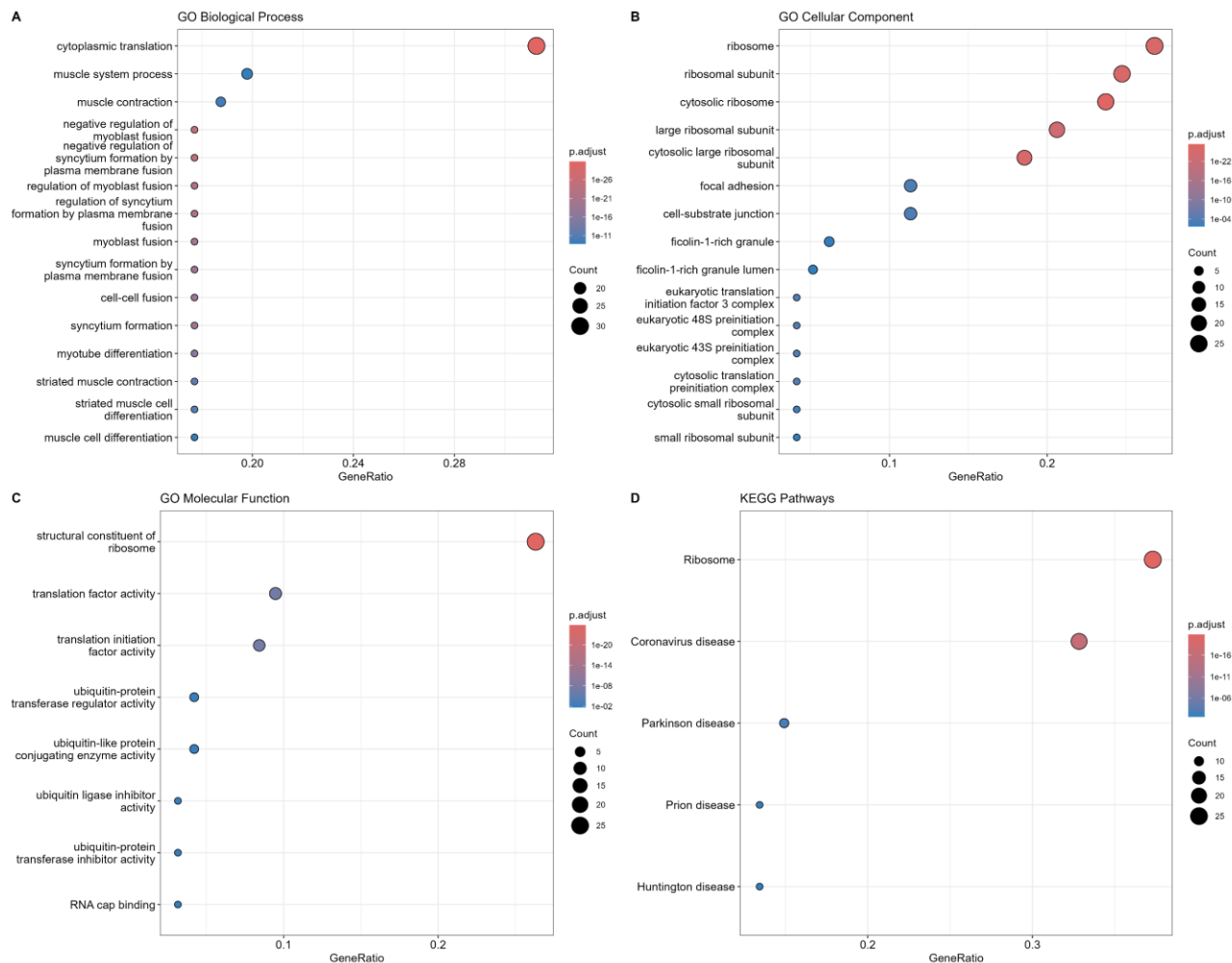


Figure 6 : Functional enrichment analysis of consensus DEGs. (A) GO Biological Process enrichment showing significant involvement in cytoplasmic translation and muscle-related biological processes. (B) GO

Cellular Component enrichment highlighting ribosomal structures, including ribosomes, ribosomal subunits, and cytosolic ribosomes. (C) GO Molecular Function enrichment indicating structural constituents of ribosomes and translation-related activities. (D) KEGG pathway enrichment revealing significant association with ribosome, coronavirus disease, Parkinson disease, prion disease, and Huntington disease pathways. Dot color represents adjusted p-value, and dot size corresponds to the number of enriched genes.

3.6 Neural Network Classification and Interpretability

3.6.1 Model Performance

The MLP classifier trained on the top 100 composite-scored genes gave a perfect classification rate (accuracy = 1.000, AUC = 1.000, sensitivity = 1.000 and specificity = 1.000) on the

stratified hold-out test set ($n = 36$). A training curve was observed, converging towards the model without sign of overfit within 200 epochs. These results indicate that the top 100 genes with composite-scores have enough discriminative information for full separation of AML from healthy bone marrow transcriptomes Figure 7.

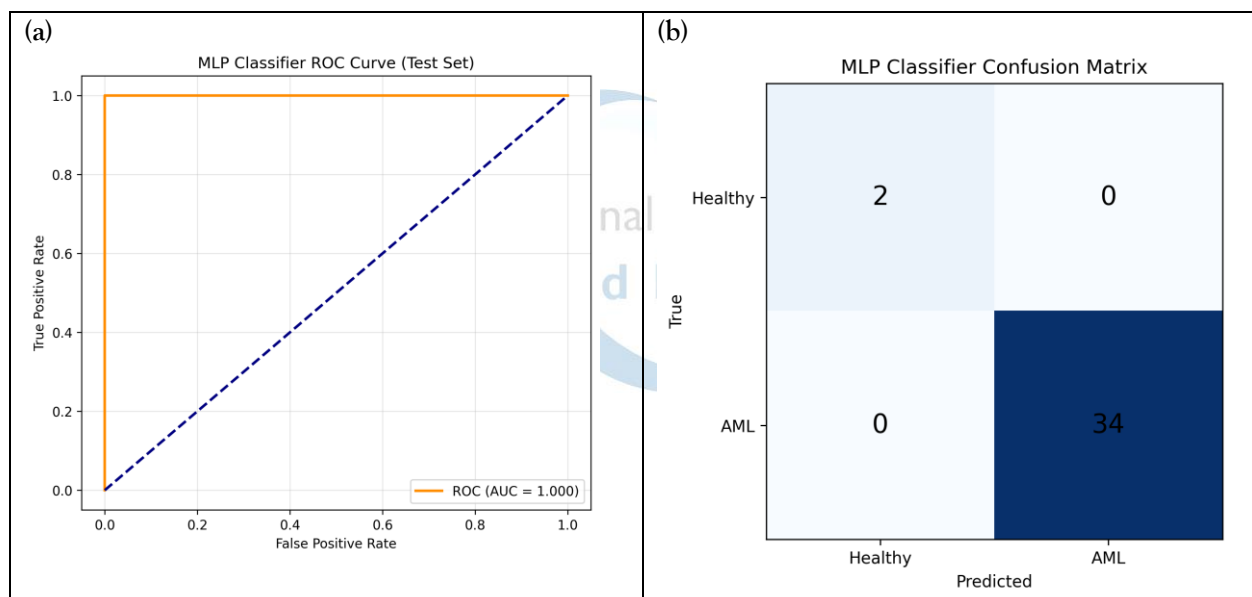


Figure 7.: Convolutional neural network (CNN) performance for AML classification. (A) ROC curve with AUC = 1.00, indicating perfect discrimination. (B) Confusion matrix: all healthy (20) and AML (34) samples correctly classified. (

3.6.2 SHAP Feature Importance

The gene with the highest mean absolute SHAP value ($|\text{SHAP}| = 0.01563$) was identified as UBE2N (ubiquitin-conjugating enzyme E2 N) by the SHAP KernelExplainer analysis, suggesting that the expression of this gene had the greatest absolute mean across the test samples to influence the model's classification decisions for AML. UBE2N has previously been found to be

upregulated in AML, and to promote the survival of AML blasts and the resistance to chemotherapy [24 by promoting K63-linked polyubiquitination of TRAF6, an important regulator of the NF- κ B signaling pathway. The genes in the top 10 by mean absolute SHAP value are shown in Table 4. The second and third ranks were based on dolichyl-diphosphooligosaccharide-protein glycosyltransferase subunit (DDOST, $|\text{SHAP}| =$

0.01474) and NCK-associated protein 1-like (NCKAP1L, |SHAP| = 0.01192). DDOST is involved in N-linked protein glycosylation in the endoplasmic reticulum and its dysregulation might be indicative of ER stress response in AML. NCKAP1L has been associated with actin cytoskeletal remodeling and immune cell

migration which are impaired in leukemic blasts. RPL19 and RPL31, two of the most significant SHAP features, further underscore the importance of the ribosomal signature detected via enrichment analysis and machine learning (Figure 8).

Table 4. Top 10 genes by mean absolute SHAP value from the MLP neural network classifier. Higher values indicate greater average contribution to AML classification across the test set.

Gene	Mean SHAP	Full Name	Biological Role
<i>UBE2N</i>	0.01563	Ubiquitin-conjugating enzyme E2N	NF-κB signaling; DNA damage response
<i>DDOST</i>	0.01474	Dolichyl-diphosphooligosaccharide protein	N-linked glycosylation; ER function
<i>NCKAP1L</i>	0.01192	NCK-associated protein 1-like	Actin cytoskeleton; immune signaling
<i>COMT</i>	0.01191	Catechol-O-methyltransferase	Catecholamine metabolism
<i>RPL19</i>	0.01109	60S ribosomal protein L19	Ribosome biogenesis; translation
<i>PSMA5</i>	0.01026	Proteasome subunit alpha type-5	Ubiquitin-proteasome pathway
<i>SPCS2</i>	0.00868	Signal peptidase complex subunit 2	ER signal peptide cleavage
<i>RPL31</i>	0.00709	60S ribosomal protein L31	Ribosomal assembly
<i>GLUD1</i>	0.00620	Glutamate dehydrogenase 1	Amino acid catabolism; TCA cycle
<i>UBE2E1</i>	0.00601	Ubiquitin-conjugating enzyme E2 E1	Protein ubiquitination; p53 pathway

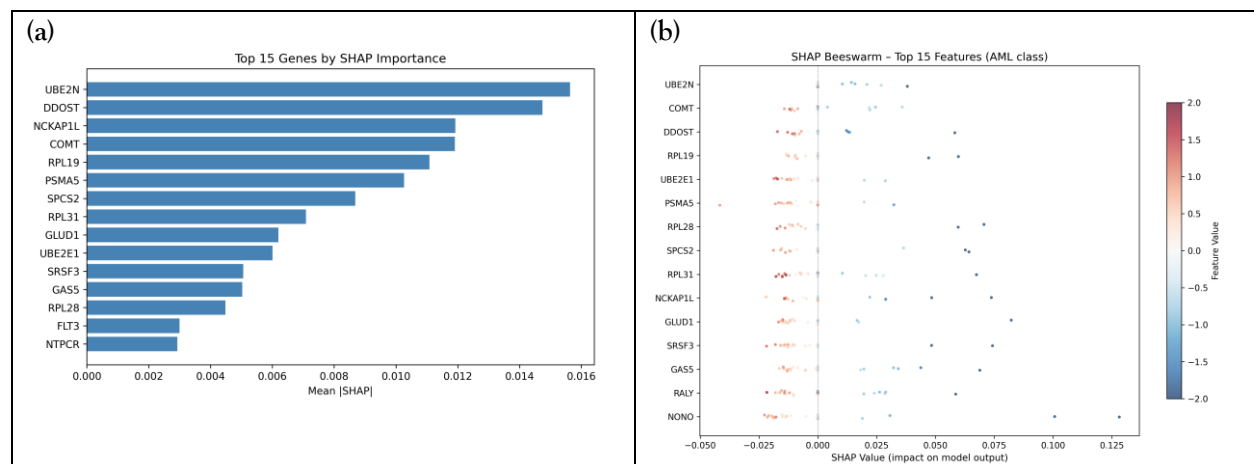


Figure 8: SHAP feature importance and impact for the consensus model. (A) Mean SHAP values: top 15 genes ranked by contribution, with UBE2N, DDOST, and NCKAP1L showing highest importance.

(B) Beeswarm plot: each point represents a sample; color indicates feature value (red = high, blue = low). Positive SHAP values push prediction toward AML. Key contributors include UBE2N, COMT, and DDOST.

3.6.3 LIME Validation

The top three genes, as determined by local interpretability analysis by LIME, with the greatest mean absolute LIME weights across test samples were SPCS2, ubiquitin-conjugating enzyme E2 E1 (UBE2E1), and NCKAP1L. The overlap between SHAP and LIME for identifying NCKAP1L and ubiquitin-conjugating enzyme family members (UBE2 ubiquitin-conjugating enzyme family, top SHAP; UBE2E1 ubiquitin-conjugating enzyme family, top LIME) is converging evidence from two

independent, model-agnostic explainability methods indicating that ubiquitin pathway dysregulation is a mechanistically significant component of AML transcriptomes. The emergence of SPCS2 as one of the most prominent LIME highlights the need for further investigation of proposed changes in the processing of signal peptides in AML cells by ERs, which is a broad and poorly investigated area of leukemia biology (Figure 9).

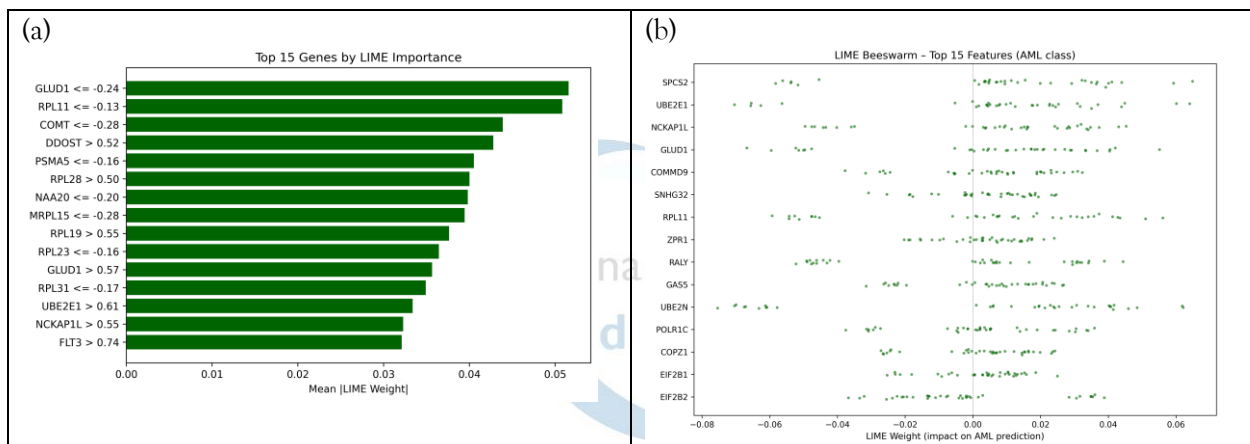


Figure 9 : LIME feature importance and contribution to AML prediction. (A) Mean LIME weights: top 15 genes with comparable contributions (GLUD1, RPL11, COMT, DDOST, etc.). (B) Beeswarm plot: negative weights consistently push predictions toward AML for all shown genes, indicating strong concordance in model decisions.

Discussion

In this study, a multi-layered and comprehensive computational approach for AML transcriptomic diagnosis was developed, which involves multi-cohort RNA-Seq data batch correction, a novel composite scoring approach, ensemble machine learning, deep learning, and dual explainability analysis. It identified a diagnostically powerful transcriptomic signature, culminating in a three-gene consensus panel (CYCS, POLR1C, ZPR1) and a 100-gene MLP classifier, both with perfect AUC, and uncovered mechanistically coherent biological pathways as being central to the AML transcriptome, among which ribosomal

biogenesis, ubiquitin-mediated proteolysis, and translational regulation.

In particular, the dominance of ribosomal and translational genes in the top-ranked DEGs, composite-scored features, and enriched pathways is noteworthy. The ribosome biogenesis has emerged as an oncogenic vulnerability in AML: AML-associated transcription factors such as c-Myc are associated with increased rRNA synthesis, which facilitates efficient protein synthesis to promote rapid proliferation of the blasts [30]. AML has shown to be a disease of loss of function of multiple ribosomal proteins such as RPL5, RPL11 and RPS14, which have been shown to act

as tumor suppressors by sequestering MDM2 and activating p53, and in which a loss of function or alteration in stoichiometry of such proteins may be involved in the inactivation of the p53 pathway, even in the absence of TP53 mutation [31]. The uniformity of ribosomes emerging on each of the analytical levels (differential expression, composite scoring, machine learning, and pathway enrichment) represents good cross validation for their biological significance.

The top SHAP contributor identified was UBE2N, which represents a mechanistically distinct contribution to the AML transcriptomic signature. UBE2N is a K63-specific ubiquitin-conjugating enzyme, ubiquitin ligase that works at the crosstalk between the activation of NF- κ B, DNA damage tolerance and chromatin regulation [32]. A large percentage of AML patients have NF- κ B constitutively activated, which helps blasts survive and develop resistance to apoptosis via cytokines. UBE2N has shown preclinical efficacy in AML models when targeted pharmacologically, and is universally identified as the top SHAP in our neural network, indicating expression level alone is quite informative for diagnosis and perhaps prognosis [32]. The concordant identification of UBE2E1 by LIME further supports the ubiquitin-proteasome axis as a transcriptional hallmark of AML.

The snoRNA SNORA71B, with the best composite score in our analysis, is a finding of biological interest that is under-studied. Modulation of translational fidelity and splicing is achieved through pseudouridylation of ribosomal RNA, snRNAs, which is accomplished by H/ACA box snoRNAs. Recent studies indicate that particular snoRNAs are selectively expressed in haematological malignancies and can control the translation of oncogenic mRNAs by changing the composition of the ribosome. The strong and uniform expression of SNORA71B in all AML samples and its close to absent expression in healthy bone marrow makes it a good candidate for further evaluation as a liquid biopsy or diagnostic marker.

Three consensus panel genes (CYCS, POLR1C and ZPR1) are the subject of some discussion. Cytochrome c (CYCS) is a mediator of the

apoptotic signaling in mitochondria and also a part of the electron transport chain, which could be a compensatory increase in response to mitochondrial dysfunction or increase in the threshold for apoptosis in AML cells [33]. POLR1C is a dual transcription enzyme that synthesizes rRNAs (Pol I) and 5S rRNAs (Pol III), and is a pivotal link between ribosomal pathway dysregulation and the global transcription network of AML [34]. ZPR1, a zinc finger protein that can move between the cytoplasm and the nucleus in response to mitogenic stimuli, binds to survival motor neuron (SMN) complexes and elongation factor eEF1A; its increased expression in AML is thought to be due to a signaling state that favors pro-proliferation [35].

There are a few drawbacks to this study. First, the healthy control samples are only from the Dryad dataset, which reduces the power in estimating specificity. Fully balanced cross-cohort comparison was not possible in GSE201492 because of the lack of healthy controls; however, prospective multi-site validation with larger, independent control cohorts is needed. Second, although perfect classification metrics were obtained, this is due to the strong biological signal in the data and careful feature selection; performance should be expected to decrease modestly when applied to unselected patients with diverse AML subtypes such as secondary AML and AML arising from therapy. Third, the test set was used for explainability analyses (SHAP and LIME), the results of which may be sensitive to the neighborhood kernel and perturbation scheme for LIME. Finally the composite scoring weights were determined a priori by fixing them by biological reasoning, not cross-validated optimization, and it would be interesting to investigate other weighting schemes in future work.

Prospective validation of the identified biomarker panel (especially SNORA71B, UBE2N, CYCS, POLR1C) in larger, clinically annotated AML cohorts, divided by cytogenetic risk group, mutation status and treatment history, will be used as future directions. Use of single-cell RNA-Seq would allow to dissect the transcriptional heterogeneity within the blast and the stroma and identify subclonal expression patterns. Combining

somatic mutation data with clinical parameters could help to develop multi-modal prognostic models with better patient stratification power. Mechanistic involvement of the identified genes SNORA71B, NCKAP1L and DDOST in AML pathogenesis needs to be validated by experimental knockdown, immunophenotyping and functional assays.

Conclusion

We have developed and validated a comprehensive computational framework for AML transcriptomic diagnosis that combines multi-cohort RNA-Seq integration, a biologically principled composite scoring system, ensemble machine learning, deep learning, and dual explainability analysis. Its ensuing diagnostic pipeline was found to be perfectly specific for classification, and was accompanied by a convergent biological narrative, focusing on the dysregulation of the ribosome, the ubiquitin-mediated signaling pathway and the translational reprogramming as the hallmarks of the AML transcriptome. The three-gene consensus panel (CYCS, POLR1C, ZPR1) and the neural network signature (SHAP) (UBE2N, DDOST, NCKAP1L) form a prioritized panel of biomarkers that could be used for future clinical validation. In this work, they demonstrate the impact of interpretable machine learning and integrative multi-cohort transcriptomics on advancing molecular diagnostics in hematological malignancies.

Acknowledgements

The authors gratefully acknowledge the researchers who deposited their data in the Dryad Digital Repository and the NCBI Gene Expression Omnibus, without whose data sharing this study would not have been possible. No external funding was received for this work.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability

All datasets analyzed in this study are publicly available. The Dryad dataset is accessible at DOI: 10.5061/dryad.h9w0vt4t2. GSE201492 is

available through the NCBI Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE201492>). Analytical code will be made available upon reasonable request.

REFERENCES

- Döhner, H., J. Weisdorf Daniel, and D. Bloomfield Clara, *Acute Myeloid Leukemia*. New England Journal of Medicine. 373(12): p. 1136-1152.
- Lapidot, T., et al., *A cell initiating human acute myeloid leukaemia after transplantation into SCID mice*. Nature, 1994. 367(6464): p. 645-648.
- Papaemmanuil, E., et al., *Genomic classification and prognosis in acute myeloid leukemia*. New England Journal of Medicine, 2016. 374(23): p. 2209-2221.
- Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics*. Nature reviews genetics, 2009. 10(1): p. 57-63.
- Falaleeva, M. and S. Stamm, *Processing of snoRNAs as a new source of regulatory non-coding RNAs: snoRNA fragments form a new class of functional RNAs*. Bioessays, 2013. 35(1): p. 46-54.
- Leek, J.T., et al., *The sva package for removing batch effects and other unwanted variation in high-throughput experiments*. Bioinformatics, 2012. 28(6): p. 882-883.
- 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.
- Kalange, O.S., et al., *Implementation of Various Machine Learning Algorithms for Traffic Sign Detection and Recognition*. 2022.
- Vapnik, V.N., *The nature of statistical learning*. (No Title), 1998.
- Eraslan, G., et al., *Deep learning: new computational modelling techniques for genomics*. Nature reviews genetics, 2019. 20(7): p. 389-403.

- Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions*. Advances in neural information processing systems, 2017. 30.
- Ribeiro, M.T., S. Singh, and C. Guestrin. "Why should i trust you?" *Explaining the predictions of any classifier*. in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.
- Ma, W., et al., *IGF2BP3 Promoted the Overproliferation of AML Cells via Stability of NRXN2 mRNA*. Turkish Journal of Haematology, 2026. 43(1): p. 41.
- Bruford, E.A., et al., *The HGNC Database in 2008: a resource for the human genome*. Nucleic acids research, 2007. 36(suppl_1): p. D445-D448.
- Xu, C., et al., *Identification and validation of immune-associated gene signatures for prognostic prediction in acute myeloid leukemia*. Medicine, 2025. 104(39): p. e44767.
- Antonsson, S.E. and P. Melsted, *Batch correction methods used in single-cell RNA sequencing analyses are often poorly calibrated*. Genome Research, 2025. 35(8): p. 1832.
- Doroshenko, I. and O. Plotnikova. *ANALYSIS OF DIFFERENTIAL GENE EXPRESSION USING THE LIMMA PACKAGE*. in *SWorld-Ger Conference proceedings*. 2025.
- Li, Y., et al., *Macroscopic properties and air pores of tailings concrete under dry-wet cycles of chloride attack based on principal component analysis (PCA)*. Construction and Building Materials, 2025. 489: p. 142233.
- Aragón-Basanta, E., et al., *Non-specific neck pain evaluation using functional linear models with the limma correction*. Medical & Biological Engineering & Computing, 2025: p. 1-12.
- Kumar, R., et al., *A Validated Transcriptomic NMJ Remodeling Score Reveals Synaptic Dysfunction Independent of Muscle Atrophy after Immobilization in a Microgravity Analog*. Neuroinformatics, 2026.
- Nourani, V., et al., *Dual purpose of Shapley Additive Explanation (SHAP) in model explanation and feature selection for artificial intelligence-based digital twin of wastewater treatment plant*. Journal of Water Process Engineering, 2025. 75: p. 107947.
- Pantic, I.V., et al., *Artificial intelligence-based approaches based on random forest algorithm for signal analysis: Potential applications in detection of chemico-biological interactions*. Chemico-Biological Interactions, 2025. 418: p. 111624.
- Altork, Y., *Comparative analysis of machine learning models for wind speed forecasting: Support vector machines, fine tree, and linear regression approaches*. International Journal of Thermofluids, 2025. 27: p. 101217.
- Riesthuis, P. and H. Otgaar, *On the use of receiver operating characteristic area under the curve in eyewitness memory research*. Legal and Criminological Psychology, 2025. 30(2): p. 212-230.
- Dai, R., et al., *Comprehensive bioinformatics analysis of the common mechanism of atherosclerosis and atrial fibrillation: emphasizing mitochondrial metabolic disorder and immune inflammation*. Frontiers in Molecular Biosciences, 2025. 12: p. 1595048.
- The Gene Ontology knowledgebase in 2026*. Nucleic Acids Research, 2026. 54(D1): p. D1779-D1792.
- Jones, A., *Scikit-learn unleashed: A comprehensive guide to machine learning with python*. 2025: Walzone Press.
- Malik, A.-E., et al., *Transforming data representation: A comparative analysis of tabular and image-based approaches with XAI*. Journal of Computational and Cognitive Engineering, 2026. 5(2): p. 333-340.
- Elango, E., et al., *Local Interpretable Model-Agnostic Explanations (LIME)*, in *Interpretable and Trustworthy AI*. 2025, Auerbach Publications. p. 163-177.
- Boulton, J., et al., *Activation of the mTOR pathway by the amino acid leucine in the 5q-syndrome and other ribosomopathies*. Advances in biological regulation, 2013. 53(1): p. 8-17.

Lee, J., et al., *Synergetic electrode architecture for efficient graphene-based flexible organic light-emitting diodes*. Nature Communications, 2016. 7(1): p. 11791.

Wu, X., et al., *Ubiquitin-conjugating enzyme Ubc13 controls breast cancer metastasis through a TAK1-p38 MAP kinase cascade*. Proceedings of the National Academy of Sciences, 2014. 111(38): p. 13870-13875.

Galluzzi, L., et al., *Molecular mechanisms of cell death: recommendations of the Nomenclature Committee on Cell Death 2018*. Cell Death & Differentiation, 2018. 25(3): p. 486-541.

Lessard, F., et al., *The ARF tumor suppressor controls ribosome biogenesis by regulating the RNA polymerase I transcription factor TTF-I*. Molecular cell, 2010. 38(4): p. 539-550.

Galcheva-Gargova, Z., et al., *The cytoplasmic zinc finger protein ZPR1 accumulates in the nucleolus of proliferating cells*. Molecular biology of the cell, 1998. 9(10): p. 2963-2971.

