

A MULTI-METHOD MACHINE LEARNING FRAMEWORK FOR IDENTIFICATION OF A CONSENSUS TRANSCRIPTOMIC SIGNATURE IN ACUTE MYELOID LEUKEMIA

Shehzad Khalil^{*1}, Tehmina Rahim², Alveena Zehra³, Muhammad Bilal⁴, Usama Khan⁵, Abdul Muqtadir⁶, Omer Khiam⁷

^{*1,4,7}Institute of Biotechnology and Genetic Engineering, The University of Agriculture, Peshawar, Khyber Pakhtunkhwa, Pakistan

²Department of Microbiology & Biotechnology, Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences and Technology (NUST), Islamabad, Pakistan

³Department of Biotechnology, Jinnah University for Women, Karachi, Pakistan

⁵Department of Biotechnology, Faculty of Basic and applied science, International Islamic University Islamabad, Islamabad, Pakistan

⁶Department of Biotechnology, Federal Urdu University of Arts, Science and Technology, Karachi, Pakistan

^{*1}shehzadkhalil911@gmail.com, ²tehminarahim141@gmail.com, ³alveena_zehra@hotmail.com, ⁴muhammad.bilal@aup.edu.pk, ⁵usamaalamkhan016@gmail.com, ⁶abdulmuqtadironlineinternship@gmail.com, ⁷omarkhiam0342@gmail.com

DOI: <https://doi.org/10.5281/zenodo.22669641>

Keywords

acute myeloid leukemia, machine learning, transcriptomics, biomarker, feature selection, LOOCV, consensus gene signature, functional enrichment

Article History

Received: 16 May 2026

Accepted: 12 June 2026

Published: 15 August 2026

Copyright @Author

Corresponding Author: *

Shehzad Khalil

Abstract

Acute myeloid leukemia (AML) is an aggressive form of hematological malignancy characterized by the abnormal proliferation of immature myeloid precursor cells, and it is highly heterogeneous at the molecular level. Although there have been great strides in the development of genomic profiles and targeted drugs, early diagnosis and molecular classification remain significant clinical challenges. The potential of transcriptomic profiling to identify disease-associated biomarkers, enhance diagnostic accuracy and provide insight into disease biology is considerable. We created and tested a comprehensive multi-method machine learning (ML) framework to uncover a strong consensus transcriptomic signature that distinguishes AML patients from healthy individuals. Transcriptomic data (10 AML samples and 10 healthy controls) was retrieved from a publicly available repository and processed using a strict pipeline of data preprocessing, normalisation and quality control (QC). After filtering low expressed genes, 13,329 expressed genes were retained, and the top 1,000 most variable genes were selected for downstream analyses. Four supervised ML algorithms were trained and tested using Leave-One-Out Cross-Validation (LOOCV): Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, Random Forest (RF) and Support Vector Machine (SVM). Five different feature-selection methods were used to increase robustness and decrease algorithm-specific bias: LASSO, Elastic Net, RF importance ranking, SVM-Recursive Feature Elimination, and Boruta. All four classifiers had an area under the receiver

operating characteristic curve (AUC) of 1.0, indicating that they could perfectly separate AML and healthy samples. A combination of several feature-selection methods resulted in an 11-gene consensus signature that included *ARRB1*, *CD44*, *FAM178B*, *HDHD5*, *HOMER3*, *MRPS23*, *NRROS*, *NTPCR*, *SNHG32*, *TCL1A* and *TCN2*. Nine of these genes showed 100% individual diagnostic accuracy; the multivariable logistic regression model had an AUC of 1.0, with 100% sensitivity, and 100% specificity. Biological processes and pathways important for AML pathogenesis, such as MAPK signaling, regulation of apoptosis, cytokine receptor activity, HIF-1 signaling, intracellular signal transduction, and differentiation of hematopoietic cells, were observed to be significantly involved in functional enrichment analysis. Together, these results demonstrate the potential of multi-method ML algorithms for identifying robust transcriptomic biomarkers in AML. The proposed 11-gene signature represents a promising approach for molecular-based diagnostic and may indicate biologically relevant targets that should be further validated in large independent cohorts and functional experimental studies.

INTRODUCTION

AML is a genetically and clinically diverse form of hematological malignancy, arising from the malignant transformation of hematopoietic stem and progenitor cells [1]. The disease is defined as uncontrolled proliferation of immature myeloid blasts in the bone marrow, peripheral blood and sometimes extramedullary tissues which leads to impaired hematopoiesis and progressive bone marrow failure [2]. AML is the most common form of acute leukemia among adults, and a significant fraction of leukemia-related deaths worldwide. Long-term survival is still unsatisfactory, especially for elderly patients and patients with high-risk molecular characteristics, although chemotherapy, stem cell transplantation, and targeted therapies have improved [3]. AML is a condition of biological complexity because of the accumulation of various genetic, epigenetic, and transcriptional abnormalities which are responsible for the disruption of normal hematopoietic differentiation and contribute to leukemic transformation [4]. Genomic studies on a large scale, such as those performed by The Cancer Genome Atlas (TCGA), have identified frequent mutations in genes that participate in signal transduction, regulation of epigenetics, transcription, and RNA processing [5]. *FLT3*, *NPM1*, *DNMT3A*, *IDH1*, *IDH2*, *RUNX1*,

CEBPA, *TET2*, *ASXL1*, and *TP53* are commonly mutated genes. These molecular changes are responsible for dysregulated proliferation, failure to differentiate, resistance to apoptosis, metabolic changes, and progressive evolution of the cells [6]. There is however, a considerable inter patient variability that makes the classification of the disease difficult and universal therapeutic approaches less effective. Over the last few years, next generation sequencing (NGS) technologies have greatly improved understanding of AML biology. Of these, transcriptomic profiling offers a global picture of the expression pattern of genes, and reflects the downstream effects of genomic and epigenomic changes [7]. Transcriptomic analysis differs from mutation-based analysis, which concentrates on discrete genetic events, by capturing the concerted work of signaling pathways, transcription factors, and the responses of cells to stress and microenvironmental interactions. As a result, gene-expression profiling has become a powerful resource for the discovery of biomarkers, the classification of AML, the prediction of prognosis and the identification of therapeutic targets. In the analysis of high-dimensional biological data sets, machine learning (ML) has become increasingly crucial. Machine learning (ML) [8] is becoming more and more important in the analysis of high-dimensional biological data sets.

Using ML algorithms, complex nonlinear interactions between thousands of molecular features can be identified that are not necessarily measurable with traditional statistical methods. In the field of oncology, ML is used in the diagnosis of cancer, in the classification of cancer subtypes, in the prediction of the cancer's response to treatment, in the development of a cancer survival model, and in the discovery of cancer biomarkers [9]. AML is an area where ML-based analyses have shown great potential to enhance risk stratification and to identify novel molecular signatures. However, many previously reported signatures of gene expression are not as reproducible as desired, as a result of variations in the datasets, pre-processing methods, feature selection methods, and model validation methods [10]. A significant hurdle in the discovery of transcriptomic biomarkers is the number of features measured (genes) is significantly higher than the number of samples (patients) available. In such cases, a single feature-selection technique might select signatures that are very dependent on particular algorithmic assumptions and are not generalizable [11]. For instance, regularized regression methods like LASSO will focus on the fewest variables [12], while ensemble methods, like the Random Forest, will focus on variables that are most important for classification [13]. Likewise, the recursive elimination methods and permutation based algorithms might give significantly different results if the dataset structure and sample size vary [14]. It is therefore important to use multiple features-selection methods to avoid methodological bias and over-fitting. In order to overcome these shortcomings, consensus-based methods have become more popular in the field of bioinformatics. Consensus frameworks can increase the robustness, biological interpretability, and reduce the false discoveries by combining multiple feature-selection strategies and selecting only those genes consistently selected by independent strategies. This is

especially useful in research with comparatively small sample sizes where algorithm-specific variability can make a significant difference in the results [15]. One key challenge in biomarker discovery research is the validation of the models. To assess a model's generalizability, cross-validation techniques assess the model's predictive performance on non-observed data. Leave-One-Out Cross-Validation (LOOCV) [16] is especially appropriate for small datasets, as it uses the maximum amount of training samples, while minimizing the estimation bias. LOOCV, together with careful feature selection and validation using independent classifiers, could be used to obtain a good estimate of the diagnostic performance. Biological interpretation is also crucial in interpreting the computational results into clinically relevant information, beyond mere predictive power. Functional enrichment analyses helps the identification of molecular pathways and biological processes associated with candidate biomarkers. These analyses can provide mechanistic connections between observed signatures and disease pathogenesis, which can aid in prioritizing therapeutic targets and experimental studies. In the current study, we created a thorough multi-method machine learning pipeline for discovering transcriptomic biomarkers in AML. We used four independent classification algorithms, five complementary feature-selection strategies, and functional enrichment analyses to arrive at a consensus gene signature that would enable us to effectively distinguish AML patients from healthy controls. In particular, we aimed to test the performance of classifiers with a strict LOOCV test, identify a consensus transcriptomic signature using multi-method feature selection, evaluate the diagnostic performance of individual and combined biomarkers, and explore the biological pathways of the identified genes. The goal of this integrative approach was to establish a solid molecular basis that will facilitate the diagnosis of AML and help elucidate leukemia biology.

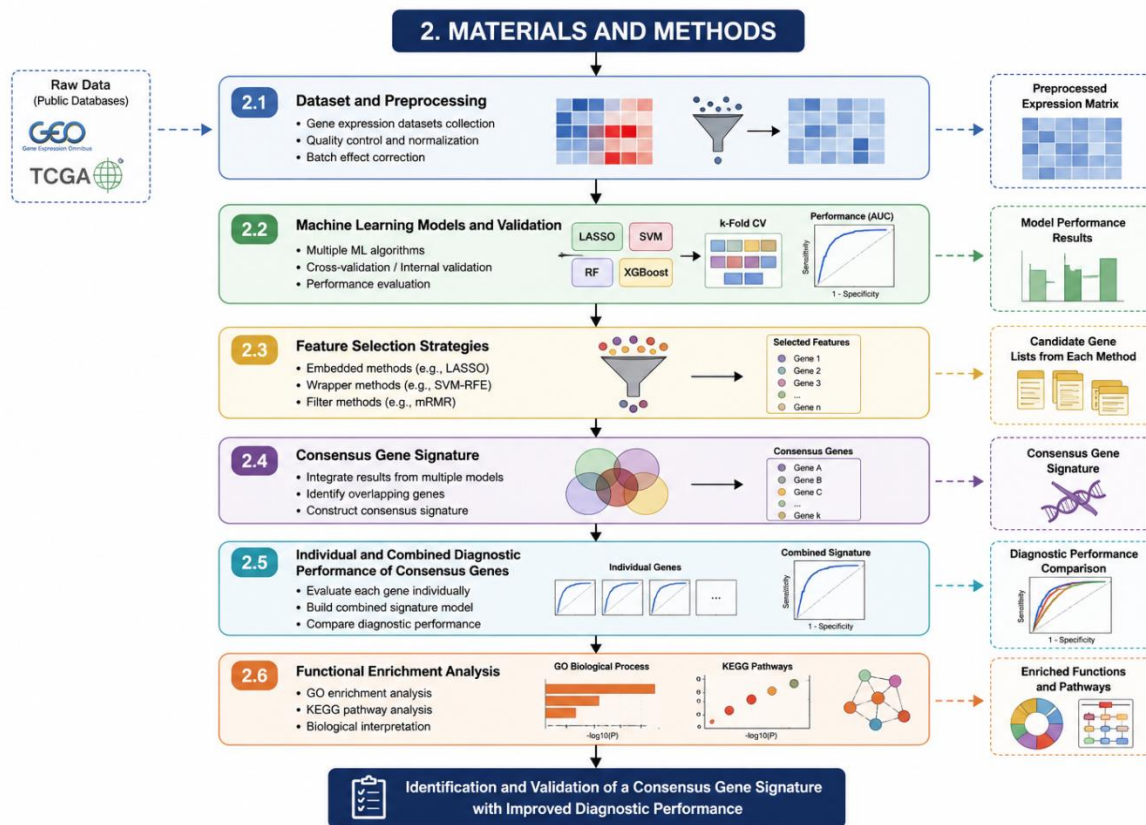


Figure 1: Schematic workflow illustrating dataset preprocessing, machine learning analysis, consensus gene signature identification, diagnostic evaluation, and functional enrichment analysis for AML biomarker discovery.

2. MATERIALS AND METHODS

Figure 1: Overview Of The Study Workflow. Gene Expression Datasets Were Preprocessed And Analyzed Using Machine Learning Models And Feature-Selection Methods To Identify A Consensus Gene Signature. The Diagnostic Performance Of Individual And Combined Genes Was Evaluated, Followed By Go And Kegg Functional Enrichment Analyses.

2.1 Dataset and Preprocessing

Transcriptomic data was retrieved from Dryad Digital Repository [17]. A dataset was created that included the gene expression data, reported in TPM, for 20 samples: 10 primary AML samples, and 10 healthy control samples. The data were retrieved in tabular format and systematically preprocessed. Version suffixes were removed from gene identifiers to promote consistency with other annotation databases. Gene-level collapsing

of the same gene symbol at the transcript level was resolved by taking the average of the duplicate genes.

TPM values were log2 transformed ($\log_2(\text{TPM} + 1)$) to normalise the data and reduce variance [18]. The genes that were in low expression in more than 80% of the samples were removed to minimize the noise created by features which were expressed near zero. The resulting expressed gene set was further filtered to top 1,000 most variable genes, the variance between the samples, which are most informative for classification tasks. All features were then Z-Score normalized (mean = 0 and SD = 1) to be comparable between the genes and models.

2.2 Machine Learning Models and Validation

Four ML classifiers were implemented and evaluated: (i) LASSO logistic regression ($\alpha = 1$) [19], (ii) Elastic Net logistic regression ($\alpha =$

0.5) [20], using a combination of L1 and L2 penalties; (iii) Random Forest (500 decision trees) [21], and (iv) Support Vector Machine (SVM) with a radial basis function (RBF) kernel [22]. The small number of individuals in the study ($n = 20$) warranted the use of Leave-One-Out Cross-Validation (LOOCV) [16] as the primary method of validation, which maximized the use of available data while giving an unbiased estimate of generalization error. For every fold in the LOOCV, one sample was randomly selected as the test set, and the rest of the 19 samples were used for the model training. Class probabilities were predicted from each fold and finally combined to create Receiver Operating Characteristic (ROC) curves, which were analysed with Area Under the Curve (AUC) calculated using pROC package in R [23]. The optimal threshold was defined as Youden index ($J = \text{sensitivity} + \text{specificity} - 1$) to obtain three performance measures: sensitivity, specificity and classification accuracy [24].

2.3 Feature Selection Strategies

Five feature selection strategies were applied to the full training dataset to identify genes most informative for distinguishing AML from healthy samples:

1. **LASSO regularization (glmnet package):** Selected 9 genes by shrinking coefficients of uninformative features to zero.
2. **Elastic Net regularization (glmnet, alpha = 0.5):** Selected 93 genes using a balanced L1/L2 penalty.
3. **Random Forest variable importance (randomForest package):** The top 10 genes ranked by Mean Decrease Gini were selected.
4. **SVM with Recursive Feature Elimination (SVM-RFE, caret and e1071 packages):** Applied with LOOCV across 1,000 features [25].
5. **Boruta algorithm (Boruta package):** A shadow-feature permutation test that identifies all relevant variables [26].

All analyses were conducted in R version 4.6.0 using the glmnet, randomForest, caret, e1071, and Boruta packages.

2.4 Consensus Gene Signature

A multi-method consensus approach was used to reduce algorithm-specific bias. Consensus genes were determined as the genes that were selected by a majority of three feature selection strategies. This threshold was selected to achieve a balance between stringency and inclusiveness, ensuring that retained genes are strongly linked with AML, across a range of algorithmic assumptions made by each individual method [27].

2.5 Individual and Combined Diagnostic Performance of Consensus Genes

Each consensus gene was evaluated for its diagnostic value alone by univariate logistic regression and AUC was calculated for the probabilities predicted by each gene versus the class labels [28]. Next, a multivariable logistic regression model was developed using all 11 consensus genes at once to evaluate the combined discriminative performance. Performance of the model was presented in terms of AUC, 95% confidence interval, sensitivity, specificity and accuracy. Confusion matrix was created to show the results of the classification.

2.6 Functional Enrichment Analysis

The biological relevance of the consensus gene signature was assessed by functional enrichment analysis with ShinyGO v0.80 [29]. The ontologies and databases queried were Gene Ontology Biological Process (GO-BP) [30], Gene Ontology Molecular Function (GO-MF) [31] and KEGG Pathways [32]. Enrichment significance was determined by Fisher's exact test with false discovery rate (FDR) correction using the Benjamini-Hochberg procedure. $FDR < 0.05$ was used as a criterion of statistical significance. Data were displayed as dot plots, where each dot represents a sample and the dots are sorted by ontology.

3. RESULTS

3.1 Dataset Characteristics and Quality Control

In the final analysis dataset, 13,329 genes were expressed in all 20 samples (10 AML, 10 controls). The most variable genes were kept for

downstream ML analysis (top 1000 most variable genes). The dataset was also balanced by having the same number of AML and control samples in the dataset, thus avoiding class-imbalance bias in the training and evaluation of the classifiers.

3.2 LOOCV Classification Performance

Under LOOCV, all four ML models—the LASSO logistic regression, Elastic Net, Random Forest, and SVM—reached perfect discrimination with an

AUC of 1.0 (Figure 2). All models had 100% sensitivity, specificity and accuracy. All four methods were completely separated at the top left corner of the ROC space, with ROC curves in all the folds coinciding. Such an excellent overall score from a variety of methodologically very different classifiers strongly indicates that the transcriptomic profiles of AML and healthy cells are very well separated in this data set.

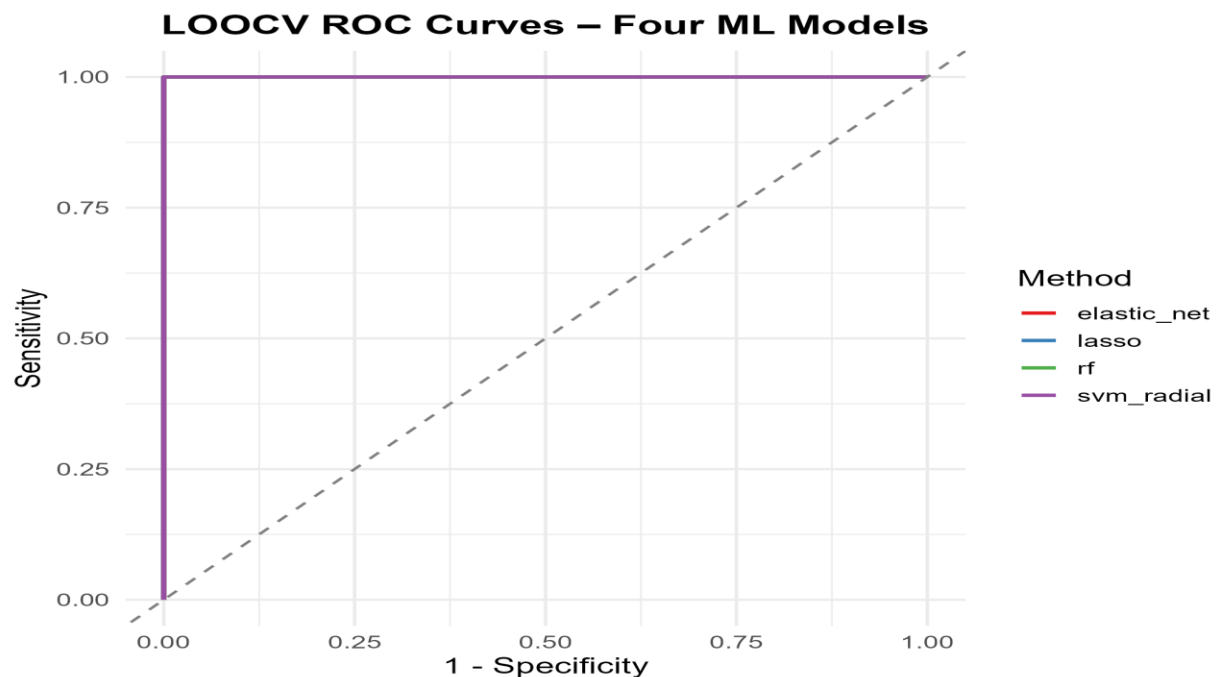


Figure 2: LOOCV ROC curves for four ML classifiers. Receiver operating characteristic (ROC) curves for LASSO, Elastic Net, Random Forest, and SVM evaluated via leave-one-out cross-validation. All models achieved AUC = 1.0.

3.3 Feature Selection and Consensus Signature

There were significantly different number of features selected for each of the five feature selection strategies (LASSO: 9 features, Elastic Net: 93 features, RF top 10: 10 features, SVM-RFE: 1000 features, and Boruta: 0 features). As is well known, in the regime of $p \gg n$, RFE fails to reduce the features, which is why SVM-RFE is applied to 20 samples and 1000 features. Using a very strict statistical threshold, Boruta did not

confirm any of the features, either. The number of features selected by each method is shown in (Figure 3, Table 1). The consensus signature comprised of 11 genes, selected by at least three methods: ARRB1, CD44, FAM178B, HDHD5, HOMER3, MRPS23, NRROS, NTPCR, SNHG32, TCL1A, and TCN2. This signature is the strong combination of regularized regression and ensemble learning methods.

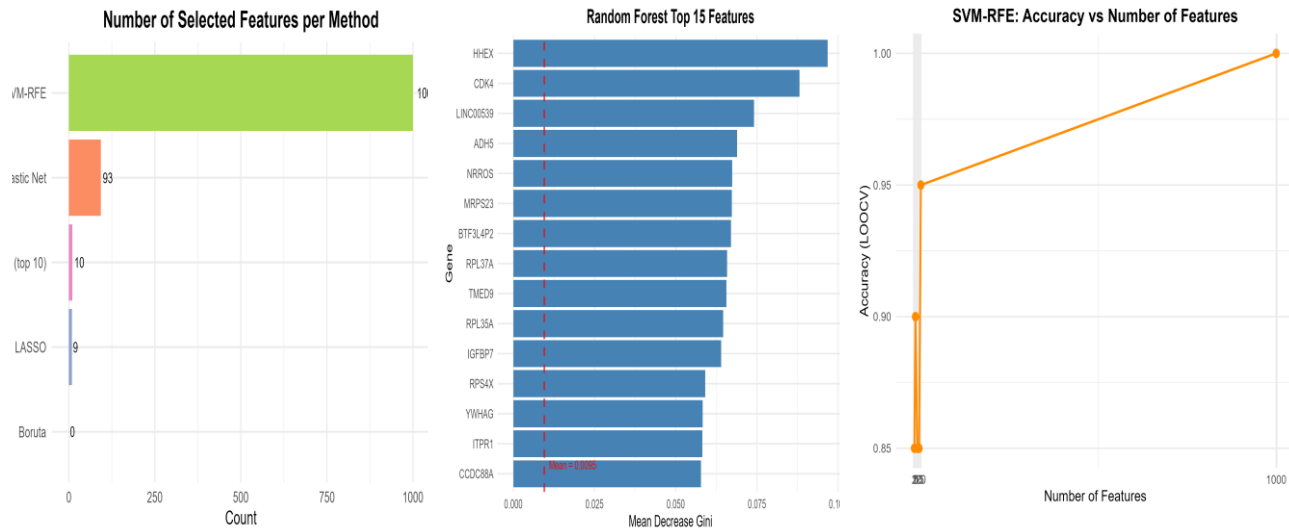


Figure 3: (a) Number of selected features per method.(b) Random Forest variable importance (top 15 genes by Mean Decrease Gini). The red dashed line indicates the mean importance across all 1,000 genes and (c) SVM-RFE accuracy across different feature subset sizes under LOOCV

Table 1: Feature selection methods and number of identified genes.

Method	Genes Selected	Approach
LASSO	9	L1 Regularization
Elastic Net	93	L1+L2 Regularization
RF (top 10)	10	Ensemble Importance
SVM-RFE	1,000	Recursive Elimination
Boruta	0	Shadow Feature Testing

3.4 Individual Diagnostic Performance of Consensus Genes

Univariate logistic regression analysis identified 9 genes, which had perfect individual discrimination between AML and healthy samples (AUC = 1.0) (Table 2, Figure 4A). Two genes—FAM178B (AUC = 0.140) and TCL1A (AUC = 0.135)—showed AUC values below 0.5. The negative AUC values suggest that, in AML

compared to healthy controls, these genes are probably downregulated and that the discriminatory potential should be assessed accordingly (e.g., an inverted probability threshold would be perfect discriminatory potential). All consensus genes have a multi-ROC curve plotted in Figure 4B which demonstrates the wide range of discriminatory space within the signature.

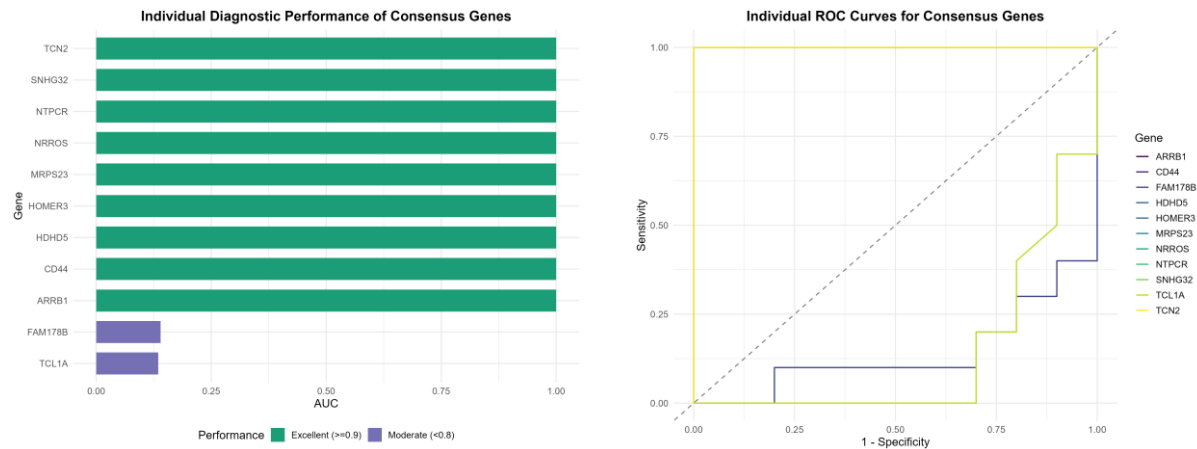


Figure 4: Individual AUC values for consensus genes, color-coded by performance category, (b) Multi-ROC curves for all consensus genes showing individual discriminatory capacity.

Table 2. Consensus genes and their individual univariate AUC values.

Gene	AUC	Performance Category
ARRB1	1.000	Excellent (≥ 0.9)
CD44	1.000	Excellent (≥ 0.9)
HDHD5	1.000	Excellent (≥ 0.9)
HOMER3	1.000	Excellent (≥ 0.9)
MRPS23	1.000	Excellent (≥ 0.9)
NRROS	1.000	Excellent (≥ 0.9)
NTPCR	1.000	Excellent (≥ 0.9)
SNHG32	1.000	Excellent (≥ 0.9)
TCN2	1.000	Excellent (≥ 0.9)
FAM178B	0.140	Poor (< 0.5)
TCL1A	0.135	Poor (< 0.5)

3.5 Combined Diagnostic Performance

The multivariable logistic regression model that included all 11 consensus genes had an AUC value of 1.0 (95% CI: 1.0–1.0), sensitivity = 100%, specificity = 100%, and overall accuracy = 100% (Figure 5a). The confusion matrix showed that there was full separation between AML and healthy samples, and that no samples were misclassified, with 0 misclassified cases in 20

cases (Figure 5b). The robustness of the results in the exceptional performance is partly due to the fact that the biological effect sizes are large, and partly due to the fact that the number of samples in this dataset is not large. However, the agreement found across four separate classifiers and five different feature selection approaches gives confidence in the diagnostic power of the signature identified.

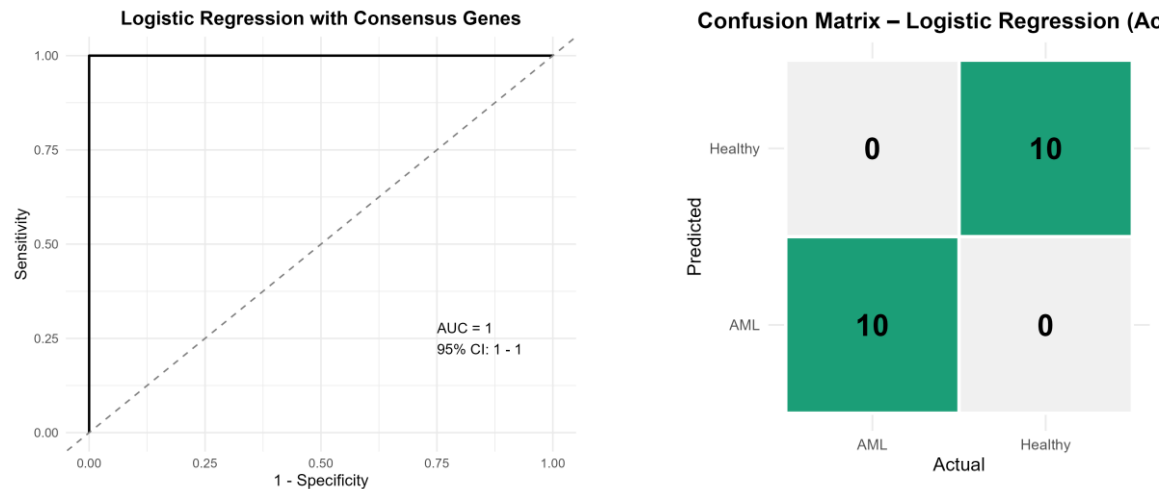


Figure 5: (a) ROC curve for logistic regression model combining all consensus genes (AUC = 1.0) (b) Confusion matrix heatmap for the logistic regression model.

3.6 Functional Enrichment Analysis

The functional enrichment analysis of the 11-gene consensus signature was performed using ShinyGO v0.80 and it was noted that there is significant overrepresentation for all three databases queried (Table 3, Figure 6). We used GO Biological Process analysis to find 381 significantly enriched terms (FDR < 0.05). The most significant biological processes were positive regulation of MAPK cascade (Fold Enrichment = 14.4, FDR = 0.027), regulation of intracellular signal transduction (Fold Enrichment = 6.2, FDR = 0.029) and regulation of apoptotic process (Fold Enrichment = 6.0, FDR = 0.042). Such processes are known to be important regulators of leukemic proliferation and survival. 36

enriched terms were identified by GO Molecular Function analysis. Cytokine receptor activity (Fold Enrichment = 37.4, FDR = 0.043) and carbon-oxygen lyase activity (Fold Enrichment = 48.3, FDR = 0.042) were the top molecular functions. Signaling through cytokines, which is a known mechanism that promotes AML blast survival and therapy resistance, is known to mediate AML blast survival and resistance to therapy [15]. The 40 pathways that were significantly enriched were mostly centered around pathway processes linked to AML pathogenesis: Hematopoietic cell lineage (Fold Enrichment = 38.5, FDR = 0.030) and HIF-1 signaling (Fold Enrichment = 35.0, FDR = 0.030) [19,7].

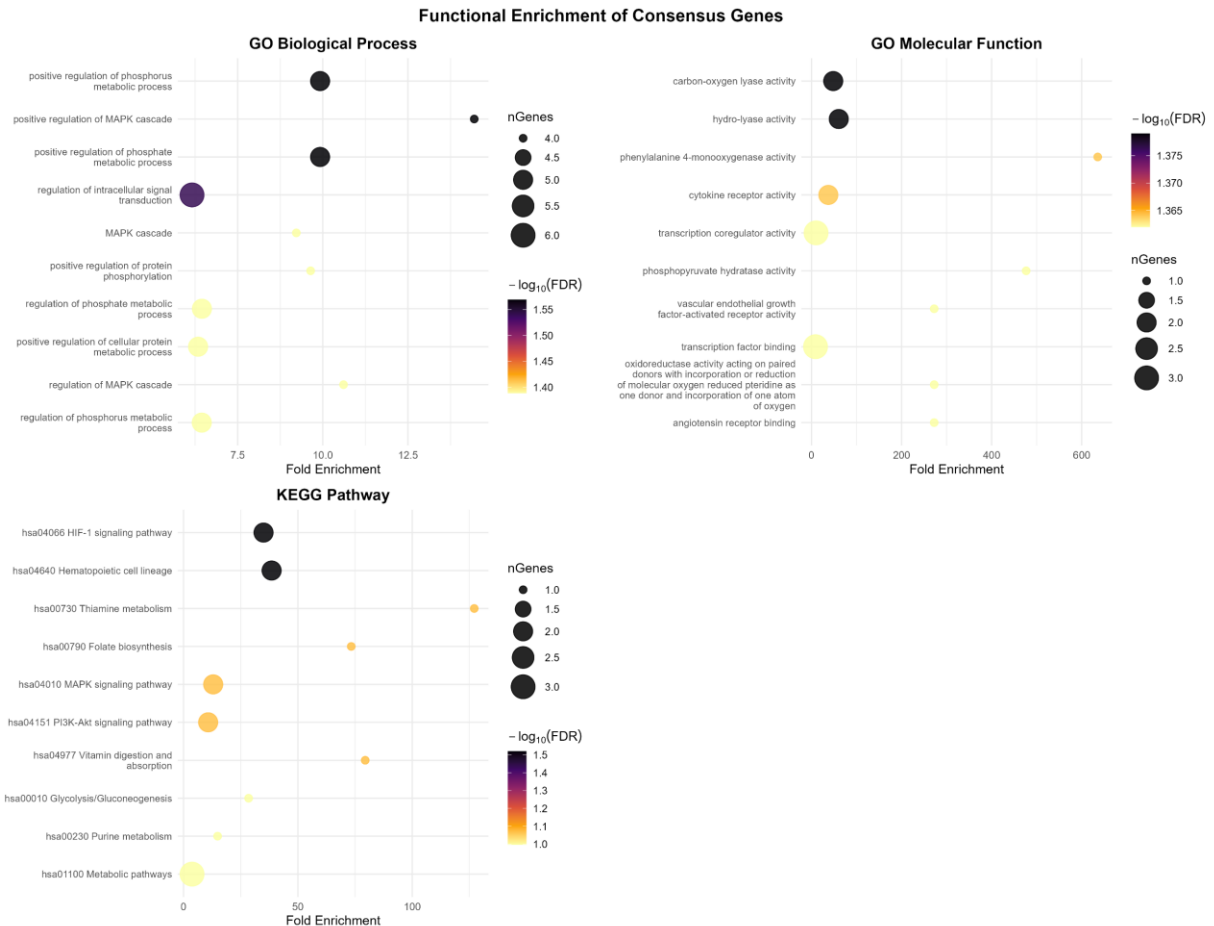


Figure 6: Combined functional enrichment dot plots for GO BP, GO MF, and KEGG pathways.

Table 3. Summary of top functional enrichment results for the consensus 11-gene signature.

Database	Top Term	Fold Enrichment	FDR
GO BP	Positive regulation of MAPK cascade	14.4	0.027
GO BP	Regulation of intracellular signal transduction	6.2	0.029
GO BP	Regulation of apoptotic process	6.0	0.042
GO MF	Carbon-oxygen lyase activity	48.3	0.042
GO MF	Cytokine receptor activity	37.4	0.043
KEGG	Hematopoietic cell lineage	38.5	0.030
KEGG	HIF-1 signaling pathway	35.0	0.030

4. DISCUSSION

Acute myeloid leukaemia (AML) is a heterogeneous haematological malignancy with many genetic and transcriptional changes that underlie disease progression, treatment resistance and poor clinical outcome [1]. To identify a robust transcriptomic signature that can

discriminate AML patients from healthy controls, we used a multi-method machine learning (ML) framework in the present study [33]. Four classification algorithms and five complementary feature-selection strategies were combined to select an 11-gene consensus signature, ARRB1, CD44, FAM178B, HDHD5, HOMER3,

MRPS23, NRROS, NTPCR, SNHG32, TCL1A and TCN2. This signature performed with good diagnostic accuracy, and was linked to biological pathways that are known to participate in AML pathogenesis. The result of this study is the perfect classification performance of all four ML models: LASSO, Elastic Net, Random Forest and Support Vector Machine. The performance of each model reached an AUC value of 1.0 when tested using LOOCV, meaning there was no overlap between AML and healthy samples. Profound gene-expression alterations in AML have been reported in previous transcriptomic studies when compared to normal hematopoietic cells [34]. The small number of samples requires interpretation with caution but the fact that the differences in transcriptomes are detected by more than one model indicates that the differences are likely to reflect real biological signals and not be a model artifact. It is considered a strength of this study to use the consensus feature-selection strategy. The individual feature-selection methods have different results due to various algorithmic assumptions and statistical criteria. We used multiple independent approaches to identify the genes that were selected, which helped reduce the possibility of algorithm-specific bias and made the biomarkers more robust [35]. The significant differences also seen amongst Boruta, SVM-RFE and regularized regression methods also underscore the significance of combining various feature selection methods when dealing with high dimensional transcriptomic data. Of the genes identified, the relevance of CD44 to the biology of AML is well established. CD44 is strongly expressed in leukemic stem cells and plays crucial roles in cell adhesion, survival and interaction with bone marrow microenvironment [36]. Previous research showed that CD44 could be considered as a biomarker and therapeutic target, as it was able to inhibit the activity of leukemic stem cells. Another important part of the signature, ARRB1, is involved in the modulation of several intracellular signaling pathways such as MAPK and PI3K/AKT, both of which are commonly dysregulated in cancer. Leukemic cells have a higher metabolic rate, which can account

for the rise of TCN2 in transport and cellular metabolism of vitamin B12 [37]. Interestingly, the expression pattern of TCL1A and FAM178B were opposite, with AUC values < 0.5 . These results do not suggest poor discriminatory ability, but rather indicate strong discriminatory ability in the other direction, in other words less expression was seen in AML compared to controls. Other genes such as SNHG32, MRPS23, NRROS, HDHD5, HOMER3 and NTPCR are relatively under-studied in AML and could be potential novel biomarkers to be further investigated. The signature was biologically validated by functional enrichment analysis. Enriched pathways were MAPK signaling pathways, pathways involved in the regulation of apoptosis, cytokine receptor activity, HIF-1 signaling pathways, and hematopoietic cell lineage pathways. The dysregulated MAPK signaling promotes AML proliferation and survival [38] and resistance to apoptosis is a hallmark of leukemic progression and therapeutic failure. In the same way, cytokine mediated signaling helps the survival of leukemic blasts in the bone marrow microenvironment [39]. The strong HIF-1 enrichment is of special interest, since pathways linked to hypoxia play a role in leukemic stem-cell maintenance and resistance to chemotherapy. In addition, AML is characterized by the disruption of normal cell lineage differentiation and enrichment of the hematopoietic cell lineage pathway. The 11-gene signature identified here could potentially be used in the clinic as a diagnostic biomarker panel. It is a relatively small, which makes it easy to implement using targeted expression assays like quantitative PCR or targeted RNA sequencing. Furthermore, several genes in the signature might be targets for therapy, especially related to signalling and microenvironmental interactions. Although these encouraging results were obtained, a few caveats need to be noted. The main disadvantage is the limited number of samples available and may lead to overfitting despite the application of LOOCV. In addition, analysis was done on a single dataset and external validation in larger and more diverse AML cohorts is needed. The signature could be used to

assess its performance in various molecular AML subtypes and its prognostic and therapeutic significance in future studies. Transcriptomic data can be coupled with genomic, epigenomic and single-cell information to further aid biomarker discovery and our understanding of AML heterogeneity. To conclude, this study revealed an 11-gene consensus transcriptomic signature that was able to reliably separate AML from healthy controls. Signature performance was excellent and showed strong links to biological pathways known to be involved in leukemogenesis, namely the MAPK signaling pathway, apoptosis regulation, cytokine signaling, HIF-1 signaling, and hematopoietic differentiation. The findings lay the groundwork for future validation studies and validate the use of integrative machine learning for the discovery of biomarkers in AML.

5. CONCLUSION

This study demonstrates that an 11-gene consensus transcriptomic signature ARRB1, CD44, FAM178B, HDHD5, HOMER3, MRPS23, NRROS, NTPCR, SNHG32, TCL1A, and TCN2 robustly distinguishes AML from healthy hematopoietic samples. The multi-method ML framework, combining LASSO, Elastic Net, Random Forest, and SVM classifiers with five feature selection strategies under LOOCV, achieved perfect diagnostic performance. Functional enrichment analysis confirmed that the consensus signature is deeply embedded in canonical AML-relevant pathways, including MAPK signaling, apoptosis, cytokine receptor activity, HIF-1 signaling, and hematopoietic cell lineage. These findings advance the molecular characterization of AML and provide a strong foundation for future biomarker validation and mechanistic studies aimed at improving AML diagnosis and therapy.

REFERENCES

1. El Chaer, F., et al., *The global epidemiology of acute myeloid leukaemia*. *Nature Reviews Clinical Oncology*, 2026. **23**(2): p. 107-120.
2. Biswas, M., *Pediatric Hematopathology: Bone Marrow Disorders, Leukemias, Lymphomas, and Molecular Diagnostics*. Birupaksha, Pediatric Hematopathology: Bone Marrow Disorders, Leukemias, Lymphomas, and Molecular Diagnostics (September 29, 2025). Biswas, B, 2025.
3. Wei, J., Q. Feng, and Z. Lu, *Leukemia in Older Adults: Global, National, and Regional Disease Burden Patterns During 1990–2019 and Forecasted Trends to 2030*. *Medicine Advances*, 2026. **4**(1): p. 38-50.
4. Soleimani Samarkhazan, H., *Integrating multi-omics approaches in acute myeloid leukemia (AML): advancements and clinical implications*. *Clinical and Experimental Medicine*, 2025. **25**(1): p. 311.
5. Dai, J., et al., *Functional germline variants together with somatic mutations alter the integrity of cancer hallmark regulatory networks*. *Genome Medicine*, 2026. **18**(1): p. 59.
6. Peng, T.-y., et al., *The role of lactate metabolism and lactylation in pulmonary arterial hypertension*. *Respiratory research*, 2025. **26**(1): p. 99.
7. Shi, J., et al., *Dynamic epigenomic landscape and gene regulatory networks during embryonic development in Pacific white shrimp (*Litopenaeus vannamei*) as revealed by histone modification profiling using CUT&Tag*. *Epigenetics & Chromatin*, 2025. **18**(1): p. 50.
8. Li, J. and S. Wang, *Integrative analysis of epigenetic subtypes in acute myeloid Leukemia: A multi-center study combining machine learning for prognostic and therapeutic insights*. *Plos one*, 2025. **20**(5): p. e0324380.
9. Afolabi, S., et al., *A Narrative Review on Machine Learning Methods for Pre-Malignant Blood Cells Identification in Hematological Malignancies*. *International Research Journal of Oncology*, 2026. **9**(1): p. 134-150.

10. Amoako, E.K.W., et al., *Exploring the role of machine learning and deep learning in anti-money laundering (AML) strategies within US financial industry: A systematic review of implementation, effectiveness, and challenges*. Finance & Accounting Research Journal, 2025. 7(1): p. 22-36.
11. Chatterjee, N., et al., *Unveiling patterns: an exploration of machine learning techniques for unsupervised feature selection in single-cell data*. Briefings in Bioinformatics, 2026. 27(1): p. bbag006.
12. Lee, T., Y. Kong, and V.P. Singh, *The more the better or the less the better: LASSO versus random forest in forecasting seasonal precipitation for drought management*. Machine Learning: Science and Technology, 2025. 6(2): p. 025019.
13. Rahardja, U. and Q. Aini, *Analyzing player performance metrics for rank prediction in valorant using random forest: A data-driven approach to skill profiling in the metaverse*. International Journal Research on Metaverse, 2025. 2(2): p. 102-120.
14. Bulut, O., et al., *Benchmarking variants of recursive feature elimination: insights from predictive tasks in education and healthcare*. Information, 2025. 16(6): p. 476.
15. Obuchowski, N.A., et al., *Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons*. Statistical methods in medical research, 2015. 24(1): p. 68-106.
16. Khadka, N., *How Leave-one-out Cross Validation (LOOCV) improve's model performance*. Dataaspirant-A Data Science Portal For Beginners, 2023. 5: p. 542-548.
17. Mayo, C., T.J. Vision, and E.A. Hull, *The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository*. International Journal of Digital Curation, 2016. 11(1): p. 150-155.
18. Zhao, Y., et al., *TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository*. Journal of translational medicine, 2021. 19(1): p. 269.
19. Meier, L., S. Van De Geer, and P. Bühlmann, *The group lasso for logistic regression*. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2008. 70(1): p. 53-71.
20. Torang, A., P. Gupta, and D.J. Klink, *An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets*. BMC bioinformatics, 2019. 20(1): p. 433.
21. Probst, P. and A.-L. Boulesteix, *To tune or not to tune the number of trees in random forest*. Journal of Machine Learning Research, 2018. 18(181): p. 1-18.
22. Thurnhofer-Hemsi, K., et al., *Radial basis function kernel optimization for support vector machine classifiers*. arXiv preprint arXiv:2007.08233, 2020.
23. Gönen, M., *Receiver operating characteristic (ROC) curves*. SAS Users Group International (SUGI), 2006. 31: p. 210-231.
24. Ruopp, M.D., et al., *Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection*. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 2008. 50(3): p. 419-430.
25. Tsiliki, G., et al., *RRegrs: an R package for computer-aided model selection with multiple regression models*. Journal of Cheminformatics, 2015. 7(1): p. 46.
26. Kurs, M.B., W.R. Rudnicki, and M.M.B. Kurs, *Package 'boruta'*. R J, 2018. 15.
27. Fröhlich, H., *Network Based Consensus Gene Signatures for Biomarker Discovery in Breast Cancer*. PLOS ONE, 2011. 6(10): p. e25364.

28. Huang, X.-W., et al., *Expert consensus on the combined screening of genes and biomarkers for neonatal diseases*. World Journal of Pediatrics, 2026. 22(1): p. 60-77.
29. Gao, H., et al., *Functional enrichment analysis of three Alzheimer's disease genome-wide association studies identifies DAB1 as a novel candidate liability/protective gene*. Biochemical and Biophysical Research Communications, 2015. 463(4): p. 490-495.
30. Pitarch, B., et al., *A review on Gene Ontology evaluations*. Database, 2025. 2025: p. baaf058.
31. Koç, I. and G. Caetano-Anollés, *The natural history of molecular functions inferred from an extensive phylogenomic analysis of gene ontology data*. PLOS ONE, 2017. 12(5): p. e0176129.
32. Du, J., et al., *A decision analysis model for KEGG pathway analysis*. BMC Bioinformatics, 2016. 17(1): p. 407.
33. Mao, J., et al., *Applying machine learning to the pharmacokinetic modeling of cyclosporine in adult renal transplant recipients: a multi-method comparison*. Frontiers in Pharmacology, 2022. Volume 13 - 2022.
34. Hsu, C.-H., et al., *Transcriptome Profiling of Pediatric Core Binding Factor AML*. PLOS ONE, 2015. 10(9): p. e0138782.
35. Obuchowski, N.A., et al., *Quantitative imaging biomarkers: A review of statistical methods for computer algorithm comparisons*. Statistical Methods in Medical Research, 2015. 24(1): p. 68-106.
36. Zöller, M., *CD44, Hyaluronan, the Hematopoietic Stem Cell, and Leukemia-Initiating Cells*. Frontiers in Immunology, 2015. Volume 6 - 2015.
37. Zhang, T.-P., et al., *Clinical relevance of vitamin B12 level and vitamin B12 metabolic gene variation in pulmonary tuberculosis*. Frontiers in Immunology, 2022. Volume 13 - 2022.
38. Ling, Q., et al., *MAP4K1 functions as a tumor promotor and drug mediator for AML via modulation of DNA damage/repair system and MAPK pathway*. eBioMedicine, 2021. 69.
39. Shafat, M.S., et al., *The bone marrow microenvironment – Home of the leukemic blasts*. Blood Reviews, 2017. 31(5): p. 277-286.