



Interpretable Machine Learning for Identifying Minimal Gene Expression Signatures Associated with Leukemia Subtypes

Dr. Pawan Motghare

Assistant Professor, Department of Oral Medicine and Radiology, Government Dental College and Hospital, Nagpur, Email: pawan21081979motghare@gmail.com, ORCID ID: 0000-0001-8374-087X

Received Date: 18-12-2025 | Revised Date: 03-02-2026 | Accepted Date: 09-03-2026 | Published Date: 28-03-2026

ABSTRACT

Leukemias can be subdivided into distinct molecular programs, although high dimensional data in the gene-expression domain makes the identification of compact and interpretable diagnostic signatures difficult. In this study, the authors present an interpretable machine-learning workflow to discriminate B-lineage acute lymphoblastic leukemia (B-ALL), T-cell ALL (T-ALL) and AML using the public expression-profiling dataset GSE7186. After quality control, the analysis was performed on 121 diagnostic samples and 5,000 expression features. Leakage-safe SelectKBest filtering was integrated in repeated stratified five-fold cross validation, along with L1 logistic regression, linear support vector machines and Extra Trees, across panels of 5, 10, 20, 50 and 100 features. The highest macro-F1 score (0.681) was obtained by L1 logistic regression using 100 features, which had a balanced accuracy of 0.690 and accuracy of 0.808. A set of 65 recurrent expression probes were identified by a stability analysis, of which 11 were present in all resampling folds. AML and T-ALL were more similar and B-ALL was more easily classified. Results reveal that compact stable transcriptomic signatures can maintain meaningful subtype-discrimination capacity while enhancing the transparency of the model and the interpretation in a biology-oriented leukemia stratification.

Keywords: Leukemia, Gene Expression, Machine Learning, Feature Selection, Molecular Classification, Bioinformatics.

1. Introduction

The leukemias form a group of biologically heterogeneous diseases of the blood, which are now being increasingly classified by combining morphology, immunophenotype, cytogenetics and molecular data. The focus on contemporary classifications is on molecularly defined disease entities since diagnosis, prognosis and therapeutic choices are affected by the differences at the genomic and transcriptomic levels. The fifth World Health Organization classification of lymphoid neoplasms establishes molecular parameters as a basis, and emphasizes the need for biologically-based typing of subtypes (Alaggio et al., 2022). There has been a parallel shift in the classification of myeloid diseases to genomically defined disease and actionable biomarkers (Khoury et al., 2022). The International Consensus Classification (ICC) is another classification that combines morphologic, clinical and genomic information in acute leukemias and related myeloid neoplasms (Arber et al., 2022).

Gene-expression profiling can capture the transcriptional state of leukemic cells in a system level representation and can capture biology associated with the various subtypes that may not be summarized by the various mutations. The GSE7186 study demonstrated that the profiles of expression of global mRNAs of childhood ALLs can be used to distinguish B-lineage ALL, T-cell ALL and AML and to fine class the child ALLs into clinically relevant molecular classes (Andersson et al., 2007). The series of corresponding Gene Expression Omnibus (GEO) data is available to support re-analysis of the transcriptomic structure associated with the leukemia subtypes, in a reproducible manner, from 121 leukemia samples and six normal bone-marrow controls (National Center for Biotechnology Information, 2007).

Machine learning is the perfect tool for high dimensional biological data where it can uncover patterns that involve multiple variables, which may not be apparent by looking at the data in a univariate manner. When the number of molecular measurements is much larger than the number of samples however, careful consideration must be given to issues of overfitting, dimensionality, validation design and interpretability for biological applications. Greener et al., (2022) stated that biologically relevant machine learning models should be developed with the focus on the scientific problem, the measuring process and the validation requirements and not just the predictive performance. Similarly, there have been examples of using machine-learning models in transcriptomic cancer research to uncover biologically meaningfully latent or selected molecular features from large-scale omics data (Malta et al., 2018).

In particular, the requirement for interpretability is important for leukemia subtype analysis. A model which can predict the subtype labels, but cannot specify which of the expression features can consistently separate the subtypes is not useful for molecular interpretation and biomarker development. However, when comparable predictive performance can be achieved, Rudin (2019) claimed that it is inherently interpretable models that are beneficial for high stakes predictive settings. Explainable AI offers complementary approaches to inspection of complex predictive models, but in the case of small-sample transcriptomic studies, perhaps the simplest approach is to use a small number of features with a sparse/linear classifier. This is a more general need that Barredo Arrieta et al. (2020) called for "opaque predictions" to be shifted toward "understandable and responsible AI systems".

This work is thus directed towards both the objectives of leukemia subtypes discrimination and signature compression. Instead of treating the entire expression space as a black-box predictor, the analysis examines increasingly smaller sub-sets of expression, quantifies the stability of selection over repeated cross-validations and compares the linear sparse methods with a nonlinear benchmark predictor, an ensemble. This design embraces the bioinformatics, genetics and genomics, molecular biology, biomedical research and human health aspects that are highlighted in the scope of CLSR. There were 3 aims of the study:

- To compare interpretable and ensemble machine-learning models for multiclass discrimination of B-ALL, T-ALL, and AML from gene-expression profiles.
- To evaluate the trade-off between predictive performance and compactness across progressively smaller expression-feature panels.
- To identify a stable consensus expression signature and characterize its reproducibility and subtype-separation structure across repeated resampling.

2. Literature Review

The most recent studies in leukemia research led by the computation of different expression-derived signatures clearly show that they can play a role in diagnosis, prognosis and molecular stratification. Hehr et al. (2023) demonstrated that an explainable AI framework can select morphologically informative cells and classify genetic AML types, demonstrating the importance of aligning the predictions of the model to clinically interpretable evidence. While their work was based on digital blood-smear morphology and not

transcriptomics, it is nonetheless a further confirmation that in order to be useful in leukemia diagnostics, AI systems need to reveal biologically or clinically meaningful decision criteria.

The direct exploration of transcriptomic feature selection in hematologic malignancies has also been attempted. Morabito et al. (2023) employed explainable deep learning to discover significant genes in chronic lymphocytic leukemia and demonstrated that the reduced set of expression feature was able to aid in clinically relevant prediction. Their findings are especially pertinent for the current study as they reveal that the gene-selection pipelines can transition from the whole-transcriptome prediction to smaller and more interpretable signatures.

In ALL, classifiers based on the expression of genes have achieved a high degree of molecular resolution. Beder et al. (2023) created ALLCatchR, an R package to classify B-cell precursor ALL molecular subtypes from RNA-sequencing expression data, demonstrating that transcriptomic patterns can capture developmental and genomic signals that define the molecular subtypes. This validates the biological assumption that a subset of expression variables can give meaningful separation of the subtypes even in the case of a complex molecular architecture.

Machine learning has also been employed to discover biomarkers associated with leukemia from more general representations of the genome. Zhang et al. (2021) used a machine-learning based stemness framework in acute leukemia to find stemness-related biomarkers. The study illustrates how feature-selection procedures and disease-specific interpretation can be combined to find molecular variables of biological relevance using predictive modeling.

In AML, the expression signatures have been created for individual risk prediction. Mosquera Orgueira et al. (2021) described a machine-learning model, using gene-expression profiling, which provided prognostic information in addition to the traditional risk factors. An evaluation of the Stellae-123 signature in subsequent independent cohorts confirmed transcriptomic predictors' reproducibility in patient cohorts and sequencing settings (Mosquera Orgueira et al., 2022).

Another alternative to the large predictor sets is the compact signatures. Liu and Elcheva (2022) reported a 6-gene prognostic signature that is shared by adult and paediatric AML, highlighting the scientific interest of such dimensionality reduction of high dimensional expression profiles to a small number of recurrently informative genes. While the methods in this study are distinct from the ones in prognosis modelling, the general goal of parsimony in molecular terms, while maintaining predictive information is somewhat similar.

When data size is small and the number of variables is large, it is important that the feature selected is stable. Nogueira et al. (2018) defined stability as the reproducibility of selected features when the data are perturbed, and established a statistical framework to consider a signature to be stable if it is reproducible when the data are resampled. This criterion can be of particular value when the number of transcriptomic cohorts is small, as a high-performing panel can have low biological credibility if its variables in the panel vary significantly from one fold to the next.

Another way of interpreting a model is by using post hoc explanation methods. SHAP (Lundberg and Lee, 2017) aims to provide a uniform framework for ascribing predictions to features, and Ribeiro et al. (2016) presented local surrogates explanations for each individual prediction. While useful for complex models, compact sparse classifiers are useful when the main scientific goal is the identification of a molecular signature that is reproducible, instead of explaining the individual prediction following model fitting.

Care is needed in evaluation measures in imbalanced multi class setting as well. Chicco and Jurman (2020) showed that, when dealing with imbalanced class distribution, Matthews correlation coefficient can deliver a more informative summary when compared to the accuracy alone. However, balanced accuracy, macro-F1 and MCC are necessary in addition to overall accuracy for the relatively small T-ALL as well as AML groups in the present cohort.

3. Methodology

3.1 Dataset and Cohort Definition

The expression-profiling series GSE7186, which was provided as public data from the NCBI Gene Expression Omnibus, was used. Swegene Human 27K RAP profiled diagnostic childhood leukemia samples comprising 87 B-lineage ALL, 11 T-cell ALL and 23 AML samples and six normal bone-marrow controls. The primary analysis was a multiclass analysis, and only the 121 leukemia samples were used to address the research question of discriminating between leukemia subtypes, not leukemia versus normal. The source sample files contain normalized log2 channel-ratio expression values after filtering (fil), pin-based LOWESS normalization (lo), hospital mean-centering (me) and presence filtering (pr) as described in the GEO records.

3.2 Expression-Matrix Construction and Quality Control

GEO sample files were programmatically extracted from compressed files without manual extraction. The identifier for each probe was matched across samples to create a matrix of samples vs. features. The features where more than 5% of the values were missing were dropped, the limited missing values were median imputed without information from the class labels and zero-variance features were excluded. To minimize computational load prior to supervised leakage, a label-free variance filter removed expression features that were not the most variable, leaving 5,000 expression features for model evaluation. All supervised feature selection was then done on the specific training sets of the cross-validation folds only.

3.3 Leakage-Safe Feature Selection and Gene-Panel Design

One-way ANOVA F statistics were used for supervised feature selection using SelectKBest. To avoid information in the test fold from affecting the feature ranking, selection was carried out within each training fold. Five different sizes of candidate panels were tested: 5, 10, 20, 50, and 100 expression features. Besides the performance in prediction, feature-selection stability was also measured by the percentage of each probe included in the repeated cross validation folds. In at least 50% of folds, probes that were selected were considered the stable consensus signature. 65 stable probes were obtained in this procedure and the highest performing predictive panel was the 100-feature configuration.

3.4 Machine-Learning Models and Cross-Validation

The comparison was made between three classifiers: L1-regularized logistic regression (one-versus-rest), linear support vector machine (SVM) and Extra Trees. Feature selection was added to the pipeline and then followed by standardization. Repeated stratified 5-fold cross validation with 5 repeats was used for model performance estimation, which resulted in 25 test folds maintaining the proportions of the different subtypes to the extent possible. The accuracy, the balanced accuracy, the macro-F1 and the MCC were computed. Macro-F1 was applied as the main criterion for comparing the feature-panel configurations as it treats all the subtypes equally regardless of the fact that there is significant class imbalance.

3.5 Signature Stability and Exploratory Molecular Structure

The feature selections were captured for all 25 resampling folds for the selected predictive configuration. The probes were ranked by selection frequency; 1.00 means that the probe has been selected in all folds. The stable consensus signature was simply used as an exploratory visualization of the expression-space structure for the principal-component analysis and was not used to estimate the predictive performance. All reported performance values are thus purely the result of repeated cross-validation.

4. Results

4.1 Cohort Composition and Class Imbalance

The final analytical cohort comprised 121 leukemia samples: 87 B-ALL (71.90%), 23 AML (19.01%), and 11 T-ALL (9.09%). The imbalance was extensive, with almost 75% of the patients having B-ALL. After performing quality control and variance filtering without supervision, 5000 expression features were kept for model development without leakage. The cohort distribution is displayed in figure 1 and the final analytical profile summarized in table 1.

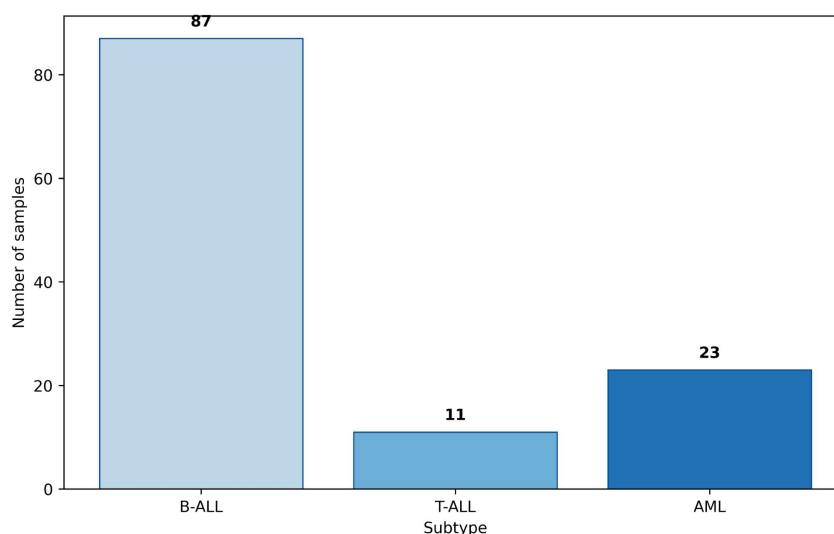


Figure 1. Leukemia subtype distribution.

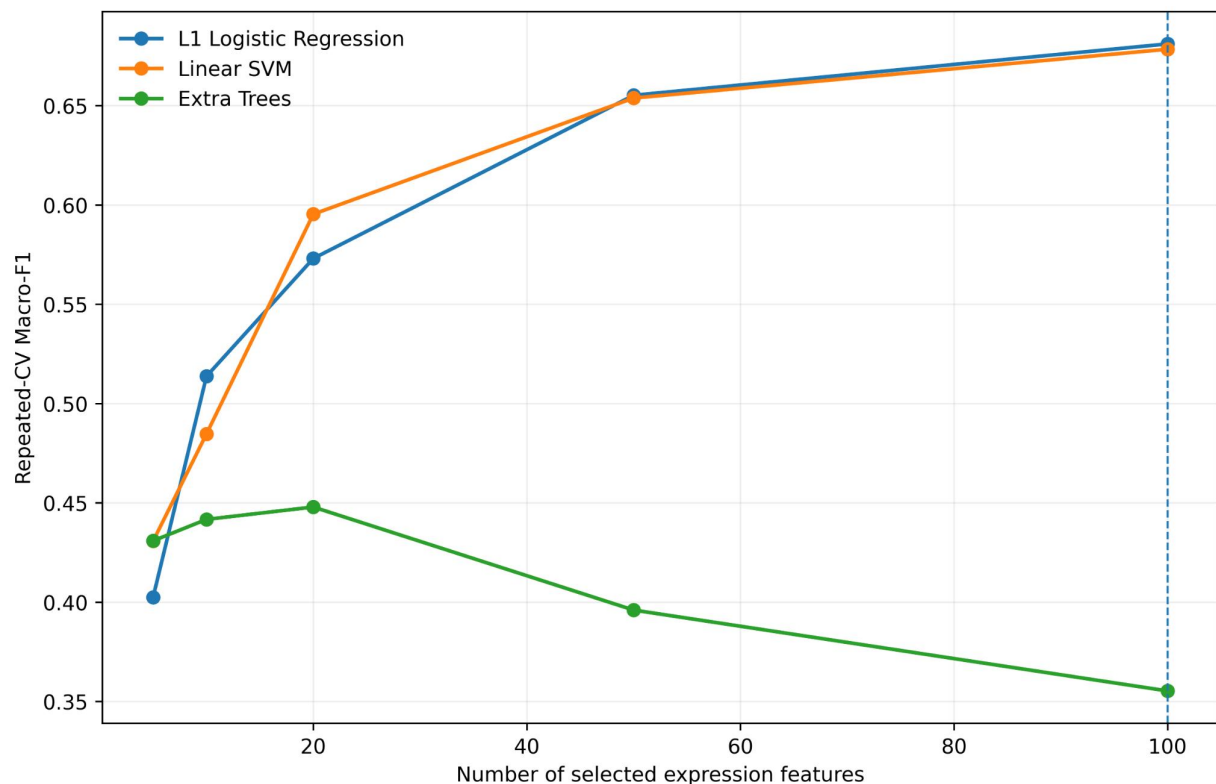
Table 1. Cohort and subtype profile.

Subtype	N	Percentage
AML	23	19.01
B-ALL	87	71.90
T-ALL	11	9.09
TOTAL_SAMPLES	121	
TOTAL_FEATURES_AFTER_QC	5000	
FINAL_MINIMAL_SIGNATURE_SIZE	65	

The minority representation of T-ALL and AML motivated the use of balanced accuracy and macro-F1 alongside overall accuracy. The table also records the 5,000-feature QC space and the 65-probe stable consensus signature identified through repeated selection.

4.2 Model Performance Across Gene-Panel Sizes

For the two linear models, performance significantly improved when the feature panel was expanded from 5 features to 50 features. L1 logistic regression improved from 0.402 (5 features) to 0.655 (50 features) to 0.681 (100 features). Linear SVM exhibited roughly the same behavior, achieving a macro-F1 score of 0.678 and with 100 features. Extra Trees, on the other hand, had its best macro-F1 at 20 features (0.448), and dropped to 100 features (0.355). These trajectories are illustrated in Figure 2, and the entire repeated-CV comparison is shown in Table 2.

**Figure 2. Repeated-CV macro-F1 across minimal gene-panel sizes****Table 2. Model and gene-panel comparison.**

Panel Size	Model	Accuracy_Mean	Accuracy_SD	Balanced_Accuracy_Mean	Balanced_Accuracy_SD	Macro_F1_Mean	Macro_F1_SD
100	L1	0.8083	0.0685	0.6895	0.1287	0.6811	0.1296

Panel_Size	Model	Accuracy_Mean	Accuracy_SD	Balanced_Accuracy_Mean	Balanced_Accuracy_SD	Macro_F1_Mean	Macro_F1_SD
	Logistic Regression						
100	Linear SVM	0.8267	0.0611	0.6888	0.1297	0.6784	0.1327
50	L1 Logistic Regression	0.7885	0.0895	0.6772	0.1387	0.6552	0.1470
50	Linear SVM	0.7884	0.0734	0.6620	0.1229	0.6538	0.1216
20	Linear SVM	0.7673	0.0952	0.6114	0.1605	0.5954	0.1586
20	L1 Logistic Regression	0.7124	0.1021	0.6196	0.1594	0.5730	0.1430
10	L1 Logistic Regression	0.6501	0.1032	0.5513	0.1663	0.5137	0.1415
10	Linear SVM	0.6914	0.0943	0.4995	0.1713	0.4847	0.1587
20	Extra Trees	0.7607	0.0611	0.4535	0.0941	0.4479	0.1251
10	Extra Tree	0.7507	0.0529	0.4471	0.0806	0.4417	0.1056

Panel_Size	Model	Accuracy_Mean	Accuracy_SD	Balanced_Accuracy_Mean	Balanced_Accuracy_SD	Macro_F1_Mean	Macro_F1_SD
	s						
5	Extra Trees	0.7473	0.0492	0.4454	0.0839	0.4308	0.0994
5	Linear SVM	0.6980	0.0674	0.4494	0.1223	0.4307	0.1174
5	L1 Logistic Regression	0.5922	0.1063	0.4163	0.1382	0.4024	0.1247
50	Extra Trees	0.7425	0.0532	0.4123	0.0745	0.3960	0.1018
100	Extra Trees	0.7244	0.0488	0.3777	0.0497	0.3553	0.0719

The best macro-F1 was achieved by the 100-feature L1 logistic regression configuration (0.681 ± 0.130), with balanced accuracy of 0.690 ± 0.129 and accuracy of 0.808 ± 0.068 . Linear SVM produced slightly higher overall accuracy at the same panel size (0.827 ± 0.061) but a marginally lower macro-F1 (0.678 ± 0.133). The 50-feature L1 model retained most of the discriminative performance, reaching macro-F1 0.655, indicating diminishing returns beyond 50 features.

4.3 Subtype-Level Classification Performance

There were significant subtype-specific recognition differences in the normalized confusion matrix. B-ALL exhibited the best classification with 90% of B-ALL observations being correctly classified in the aggregated repeated-CV test predictions. A correct classification rate of 62% was obtained for T-ALL, 15% of cases were classified as B-ALL and 24% as AML. AML had 56% correct classification and most commonly was misclassified as B-ALL (33%). These class specific patterns are illustrated in figure 3.

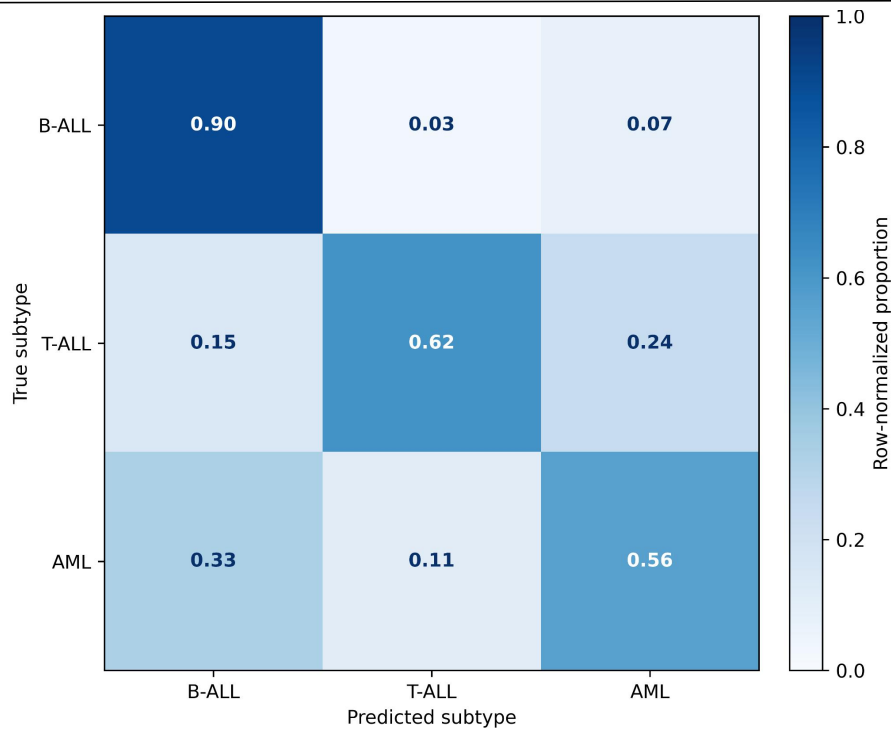


Figure 3. Repeated-CV normalized confusion matrix for the selected L1 logistic-regression configuration.

Table 4. Repeated-CV performance of the final predictive configuration.

Metric	Mean	SD
Accuracy	0.8083	0.0699
Balanced Accuracy	0.6895	0.1313
Macro F1	0.6811	0.1323
MCC	0.5723	0.1429

Across the 25 repeated test folds, the final predictive configuration achieved mean accuracy of 0.808, balanced accuracy of 0.690, macro-F1 of 0.681, and MCC of 0.572. The lower balanced accuracy relative to overall accuracy reflects the strong cohort imbalance and the greater classification difficulty of AML and T-ALL.

4.4 Stability of the Expression Signature

Selection stability showed that a core set of probes was highly reproducible across resampling. Eleven probes—18956, 18963, 3864, 6361, 6369, 6511, 6738, 6740, 8438, 8439, and 8975—were selected in all 25 folds. Three additional probes (10259, 6526, and 6554) were selected in 96% of folds, and four probes (10670, 12734, 3073, and 7306) were selected in 92% of folds. Overall, 65 probes met the predefined stability threshold of at least 50% selection frequency. Table 3 reports the ranked stability profile.

Table 3. Leading probes in the minimal-signature stability analysis (top 30 shown; the complete ranked list is retained in the analysis output).

Feature	Selection_Count	Selection_Frequency	Selected_in_Final_Signature
18956	25	1	True
18963	25	1	True
3864	25	1	True
6361	25	1	True

Feature	Selection_Count	Selection_Frequency	Selected_in_Final_Signature
6369	25	1	True
6511	25	1	True
6738	25	1	True
6740	25	1	True
8438	25	1	True
8439	25	1	True
8975	25	1	True
10259	24	0.9600	True
6526	24	0.9600	True
6554	24	0.9600	True
10670	23	0.9200	True
12734	23	0.9200	True
3073	23	0.9200	True
7306	23	0.9200	True
10941	22	0.8800	True
13465	22	0.8800	True
10276	21	0.8400	True
10442	21	0.8400	True
18046	21	0.8400	True
9573	21	0.8400	True
9667	21	0.8400	True
10939	20	0.8000	True
2193	20	0.8000	True
4430	20	0.8000	True
9708	20	0.8000	True
12642	19	0.7600	True

The repeated selection of multiple probes across all resampling iterations suggests that subtype discrimination was not driven by a single unstable training split. At the same time, the presence of a broader 65-probe consensus signature indicates that multiple correlated transcriptional signals may contribute to leukemia subtype separation.

4.5 PCA Structure of the Stable Consensus Signature

Molecular structure by principal-component analysis of the stable consensus signature gave a view of the molecular structure associated with the subtypes. The first component accounted for 73.5% of the variance and the second component accounted for 4.8% of the variance (78.3% for the two components combined). The B-ALL samples were clustered together along a relatively tight spectrum of negative and central PC1

values. Overall, AML samples appeared to be more likely to lie in higher values on PC1 but there was significant overlap between AML and B-ALL. The samples of T-ALL were more widely spread, likely due to the small number of samples in this class as well as the low-class recognition rate in the confusion matrix. Figure 4 thus corroborates the notion of transcriptional structure related to subtypes, but is not indicative of complete unsupervised separation.

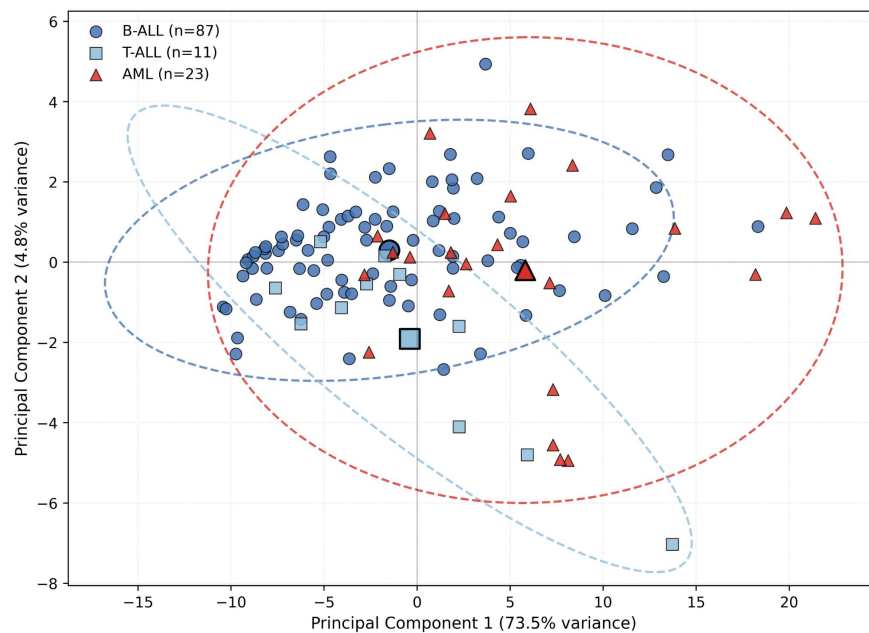


Figure 4. Leukemia subtype separation using the stable minimal gene-expression signature

5. Discussion

An interpretable machine learning workflow is presented showing that a compact and reproducible expression signature can be recovered from a high-dimensional pediatric leukemia data set and is still useful for meaningful multiclass discrimination in a leakage-safe manner. L1 logistic regression with 100 features selected performed the best; and L1 logistic regression with 50 features selected maintained majority of macro-F1 performance. This trend suggests that the majority of predictive information is contained in a relatively small number of the 5000 QC retained expression features, and that the information gained by including other features beyond the intermediate panel sizes is increasingly small.

In this context, it is biologically and statistically feasible that this is the case, that sparse and linear methods outperform Extra Trees. The number of samples in the cohort is small (121 samples) and the feature space is still large despite variance filtering. Flexible tree ensembles might split small minority classes and overfit the unstable interactions, while linear models with embedded regularization can be used to exploit the coordinated expression shifts, but also limit the complexity of the model. This finding is in general agreement with the broader idea that model complexity should match the size of the sample and the structure of the biological measurements than that it will be better if it is nonlinear (Greener et al., 2022).

But the results for each class are also crucial. The normalized sensitivity for the recognition of B-ALL was 90%, T-ALL 62% and AML 56%. The reason for this asymmetry is partly due to the class imbalance (only 11 samples in the T-ALL class) but it could also indicate that there are common transcriptional programs across acute leukaemias. This was reflected in the current classification of disease which recognizes the molecular heterogeneity of the lymphoid and myeloid leukemia entities which can not always be distinguished by one diagnostic axis. Molecularly resolved entities are in the foreground of ALL in the WHO lymphoid framework (Alaggio et al., 2022), and recurrent genomic abnormalities are in the foreground of AML in the WHO myeloid framework (Khouri et al., 2022).

The stability analysis brings an interpretability aspect which would not be manifested by the aggregate predictive metrics. Eleven probes were chosen in each resampling fold and 65 probes were above the 50% stability threshold. This repetition is significant since instability of features is a well-recognized problem with high-dimensional low sample size data (Nogueira et al., 2018). A signature which is reproduced repeatedly under the perturbation is more defensible as a candidate molecular panel than a one time ranking based on the whole data set.

The present observations also relate to the ongoing research in the areas of explainable and expression-

based leukemia modeling. Hehr et al. (2023) showed that explainable AI can reveal diagnostically-relevant pieces of evidence from cell morphology for the discrimination of AML subtypes. Transcriptomic CLL: Morabito et al. (2023) demonstrated that explainable feature-selection algorithms can be used to identify smaller, influential sets of genes. Beder et al. (2023) also showed that gene-expression classifiers can accurately classify molecular B-cell precursor ALL (BCP-ALL) subtypes in large cohorts of patients with RNA-sequencing. As an overall trend in hematologic AI, these studies suggest that there should be a growing emphasis on the identification of concise, reproducible, and biologically interpretable molecular signals and not just opaque high dimensional prediction.

It's worth recalling the difference between the best of the best—100-feature predictive panel—and the stable consensus signature (65 probes). The repeatedCV results presented here are for the 100 feature L1 logistic-reg model, while the 65 probe-set is the intersection of the variables selected in $\geq 50\%$ of the resampling folds. As such, the consensus signature is a molecular feature set indicative of reproducibility instead of an independently validated 65-feature predictive model. The PCA with this signature showed significant overall structure and significant overlap, thus supporting the caution to not overstate the separation of subtypes.

There are some caveats to be taken into account. First, the study did not incorporate an external validation data set and re-analysed a single historical microarray cohort. Second, there are few samples of the T-ALL subtype which increases uncertainty across folds. Third, the stable features reported are those of each probe, and the mapping of these probes to gene symbols and functional pathways of today would increase the strength of the biological interpretation. Fourth, the variance prefilter and the stability threshold were selected for practical reasons and the computational and methodological limitations and might affect the signature. Lastly, molecular classification is not discussed in the analysis, but prognosis and treatment response. Future studies should be performed to validate stable probes in another RNA-sequencing or microarray cohort, to map probes to genes and pathways, to investigate the biological functions of each of the subtypes, and to determine whether a smaller panel can maintain performance in a prospective and/or external study.

6. Conclusion

In this study, we used the GSE7186 cohort to create an interpretable machine-learning framework to identify compact gene-expression signatures that are associated with major pediatric leukemia subtypes. Analysis revealed that nonlinear Extra Trees model was not suitable for the high-dimensional and small-sample transcriptomic setting, and sparse linear approaches were more appropriate. The most comprehensive performance balance was for L1-regularized logistic regression with a 100-feature panel, with a mean accuracy of 0.808, balanced accuracy of 0.690, macro-F1 of 0.681 and Matthews correlation coefficient of 0.572. Performance gradually increased with the number of features selected, but with some diminishing returns for larger expression panels (over 50 features). Stability analysis also revealed 65 recurrent probes (11 of which were selected in every repeated cross-validation fold) suggesting that there is a core expression pattern associated with the subtype that is reproducible. The misclassification was most prominent for AML and T-ALL, with a large amount of overlapping in principal-component space, possibly due to the biological and statistical difficulties of smaller groups of these subtypes. Overall, the results indicate that genome-scale expression data can be summarized into a much smaller and stable set of features, which still provide discriminatory information. The proposed framework can serve as a practical foundation for bioinformatics-based stratification of leukemia and discovery of interpretable biomarkers. However, external cohort validation, probe-to-gene annotation, functional pathway analyses and experimental validation is needed prior to considering identified expression signatures for translational or clinical applications.

References

1. Alaggio, R., Amador, C., Anagnostopoulos, I., et al. (2022). The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. *Leukemia*, 36(7), 1720–1748. <https://doi.org/10.1038/s41375-022-01620-2>
2. Andersson, A., Ritz, C., Lindgren, D., Edén, P., Lassen, C., Heldrup, J., Olofsson, T., Råde, J., Fontes, M., Porwit-MacDonald, A., Behrendtz, M., Höglund, M., Johansson, B., & Fioretos, T. (2007). Microarray-based classification of a consecutive series of 121 childhood acute leukemias: Prediction of leukemic and genetic subtype as well as of minimal residual disease status. *Leukemia*, 21, 1198–1203. <https://doi.org/10.1038/sj.leu.2404688>
3. Arber, D. A., Orazi, A., Hasserjian, R. P., et al. (2022). International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: Integrating morphologic, clinical, and genomic data. *Blood*, 140(11),

- 1200–1228. <https://doi.org/10.1182/blood.2022015850>
4. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
 5. Beder, T., Hansen, B.-T., Hartmann, A. M., et al. (2023). The gene expression classifier ALLCatchR identifies B-cell precursor ALL subtypes and underlying developmental trajectories across age. *HemaSphere*, 7(9), e939. <https://doi.org/10.1097/HS9.0000000000000939>
 6. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6. <https://doi.org/10.1186/s12864-019-6413-7>
 7. Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23, 40–55. <https://doi.org/10.1038/s41580-021-00407-0>
 8. Hehr, M., Sadafi, A., Matek, C., Lienemann, P., Pohlkamp, C., Haferlach, T., Spiekermann, K., & Marr, C. (2023). Explainable AI identifies diagnostic cells of genetic AML subtypes. *PLOS Digital Health*, 2(3), e0000187. <https://doi.org/10.1371/journal.pdig.0000187>
 9. Khoury, J. D., Solary, E., Abla, O., et al. (2022). The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia*, 36(7), 1703–1719. <https://doi.org/10.1038/s41375-022-01613-1>
 10. Liu, Z., & Elcheva, I. (2022). A six-gene prognostic signature for both adult and pediatric acute myeloid leukemia identified with machine learning. *American Journal of Translational Research*, 14(9), 6210–6221.
 11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
 12. Malta, T. M., Sokolov, A., Gentles, A. J., et al. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*, 173(2), 338–354.e15. <https://doi.org/10.1016/j.cell.2018.03.034>
 13. Morabito, F., Adornetto, C., Monti, P., et al. (2023). Genes selection using deep learning and explainable artificial intelligence for chronic lymphocytic leukemia predicting the need and time to therapy. *Frontiers in Oncology*, 13, 1198992. <https://doi.org/10.3389/fonc.2023.1198992>
 14. Mosquera Orgueira, A., Peleteiro Raíndo, A., Cid López, M., et al. (2021). Personalized survival prediction of patients with acute myeloblastic leukemia using gene expression profiling. *Frontiers in Oncology*, 11, 657191. <https://doi.org/10.3389/fonc.2021.657191>
 15. Mosquera Orgueira, A., Peleteiro Raíndo, A., Díaz Arias, J. Á., et al. (2022). Evaluation of the Stellae-123 prognostic gene expression signature in acute myeloid leukemia. *Frontiers in Oncology*, 12, 968340. <https://doi.org/10.3389/fonc.2022.968340>
 16. National Center for Biotechnology Information. (2007). GSE7186: Microarray-based classification of a consecutive series of 121 childhood acute leukemias [Data set]. *Gene Expression Omnibus*. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7186>
 17. Nogueira, S., Sechidis, K., & Brown, G. (2018). On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174), 1–54. <https://jmlr.org/papers/v18/17-514.html>
 18. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
 19. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
 20. Zhang, Y., Liu, D., Li, F., Zhao, Z., Liu, X., Gao, D., Zhang, Y., & Li, H. (2021). Identification of biomarkers for acute leukemia via machine learning-based stemness index. *Gene*, 804, 145903. <https://doi.org/10.1016/j.gene.2021.145903>