From Noise to Signal: Classifying Leukemia with Gene Expression Data

A Machine Learning Approach to Differentiating ALL and AML

The Clinical Challenge: Distinguishing Two Critical Leukemia Types

ALL (Acute Lymphoblastic Leukemia)

?

AML (Acute Myeloid Leukemia)

Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) are distinct cancers requiring different treatment protocols.

Misdiagnosis can have severe consequences.

goal: Build robust classification models to predict ALL vs. AML using microarray gene expression data.

Data Source

- **Dataset**: Gene Expression Omnibus (GEO) Dataset GSE13159
- **Initial Samples**: 750 ALL, 542 AML
- **Technology**: Affymetrix Human Genome U133A Array (GPL570)

Approach: The Data Refinery Pipeline

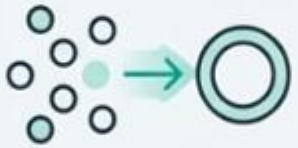
Stage 1: Ingestion

Raw GEO Data Files

Stage 2: Purification

Label Extraction

Stage 3: Refinement



Probe-to-Gene Mapping

Stage 4: Preparation

Data Splitting & Scaling

Stage 5: Quality Control

Cross-Validation & Evaluation

DATASET CONTAINED

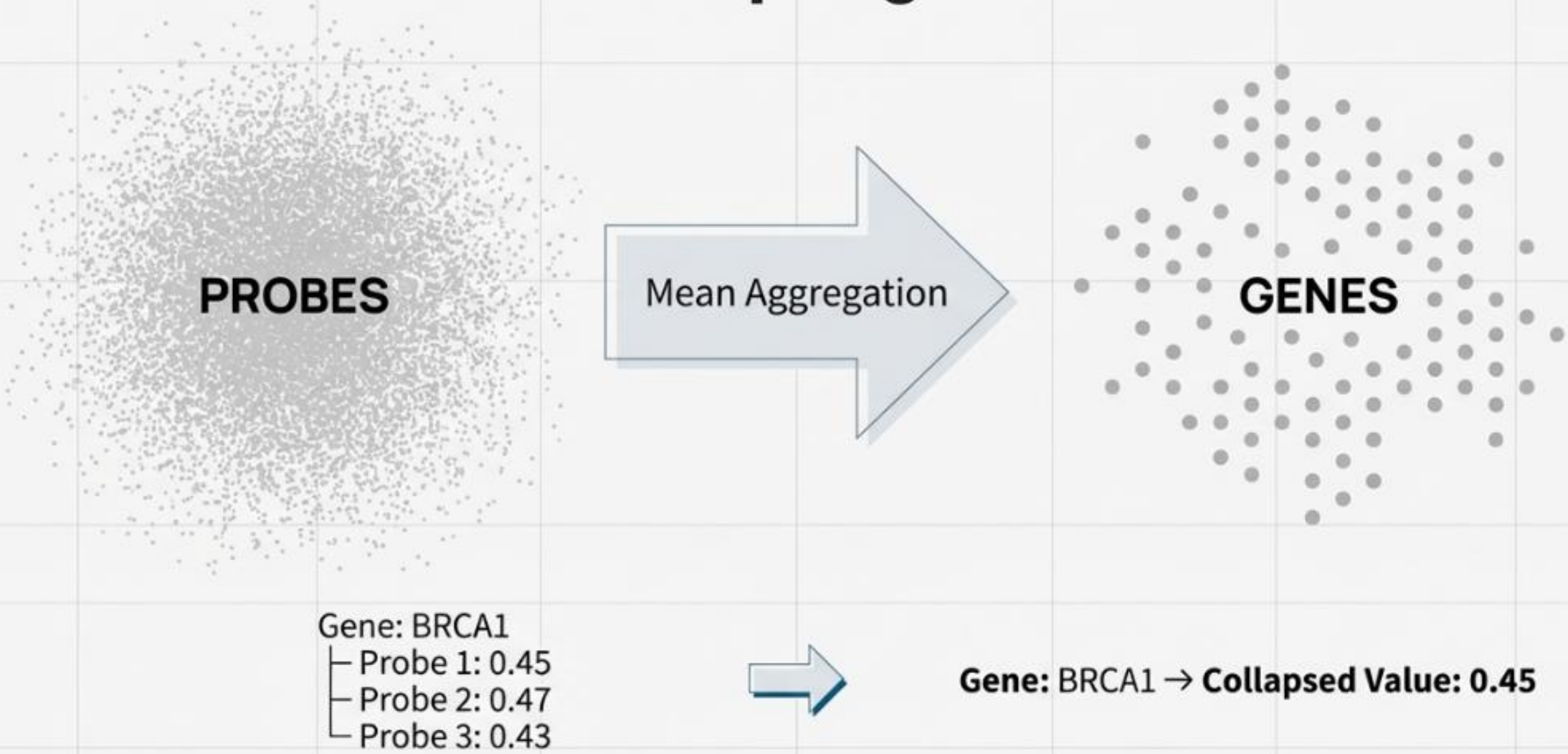
54,676 × **2,097**

Rows: Gene probes

Columns: Patient samples

Data values are normalized gene expression levels (0.0 to 1.0). Probes are short DNA sequences that measure gene activity; a single gene can have multiple probes.

The crucial breakthrough: collapsing 54,675 probes into 22,189 unique genes.



The Contenders: A Spectrum of Classifier Architectures

Linear Models

Logistic Regression, Linear SVM

Strong, interpretable baselines that excel in high-dimensional spaces, especially when classes are well-separated. SVM specifically seeks the maximum-margin hyperplane.

Tree-Based Ensembles

Random Forest, XGBoost

Powerful methods that can capture non-linear relationships. XGBoost is a state-of-the-art gradient boosting algorithm known for its performance and regularization.

Deep Learning

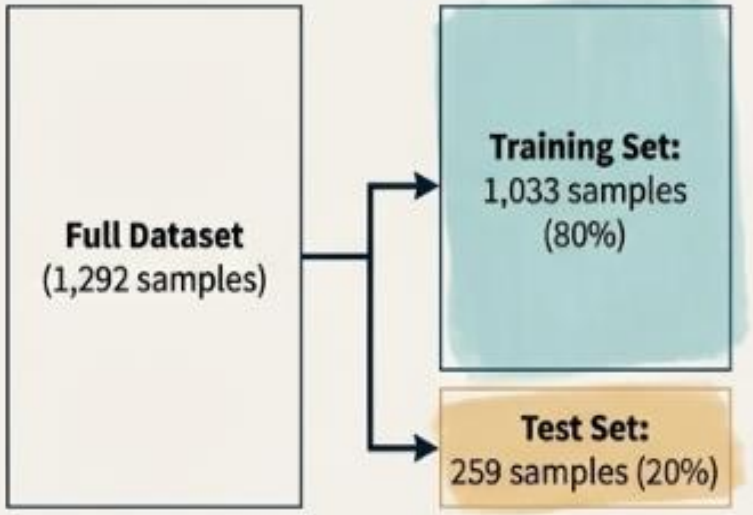
Deep Neural Network (DNN)

A modern benchmark to test if automatic feature engineering via hidden layers can discover complex biological interactions missed by other models.

Designing a Fair Evaluation Framework

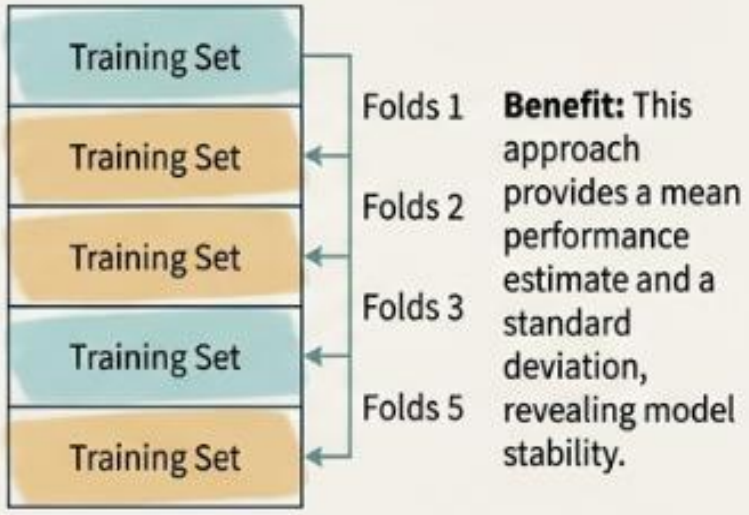
Stratified Train-Test Split

Why it Matters: dataset is imbalanced (750 ALL vs. 542 AML, a 1.38:1 ratio). Stratified splitting ensures that the training and test sets preserve this ratio, preventing biased performance estimates.



5-Fold Stratified Cross-Validation

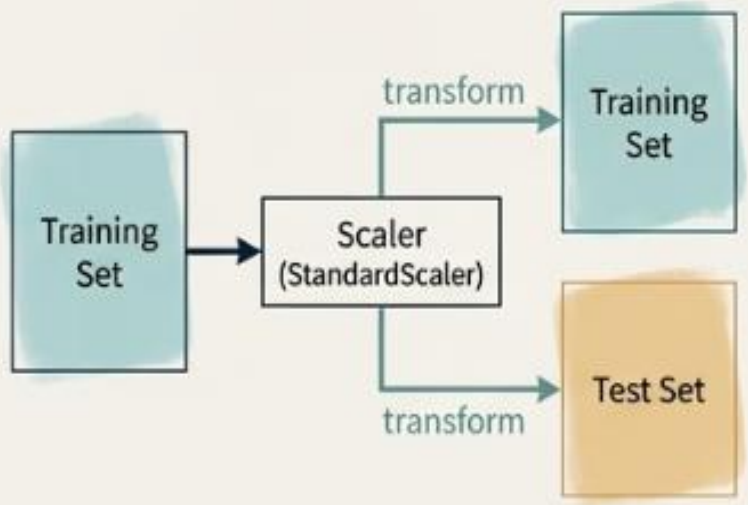
Why it Matters: To get a robust estimate of generalization performance, we average results across 5 'folds' of the training data. Stratification again ensures each fold reflects the true class balance.



Benefit: This approach provides a mean performance estimate and a standard deviation, revealing model stability.

Preventing Data Leakage

Key Principle: Feature scaling (`StandardScaler`) is critical for distance-based models. We fit the scaler **only** on the **training data** and then use it to transform the test data, avoiding any information leak from the test set into the model.



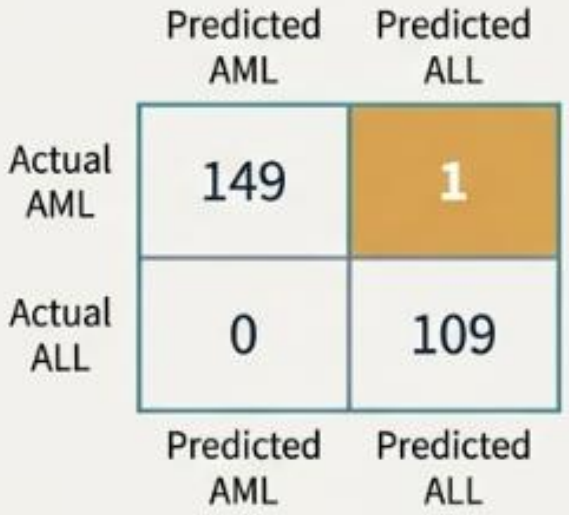
Best Overall Performance

Model Name	Accuracy	Sensitivity	Specificity	Precision	F1-score	ROC-AUC	**Key Weakness**
Logistic Regression	0.9923	1.0000	0.9867	0.9820	0.9909	1.0000	2 False Positives
Linear SVM	0.9961	1.0000	0.9933	0.9909	0.9954	0.9999	**1 False Positive**
Random Forest	0.9884	1.0000	0.9800	0.9732	0.9864	0.9995	3 False Positives
XGBoost (gene-level)	0.9923	0.9817	1.0000	1.0000	0.9907	0.9998	2 False Negatives
DNN (gene-level)	0.9614	1.0000	0.9333	0.9160	0.9561	0.9980	10 False Positives

Linear SVM achieves the highest accuracy and specificity with the lowest error rate, demonstrating that for this problem, maximizing the margin of a linear separator is the most effective and robust strategy.

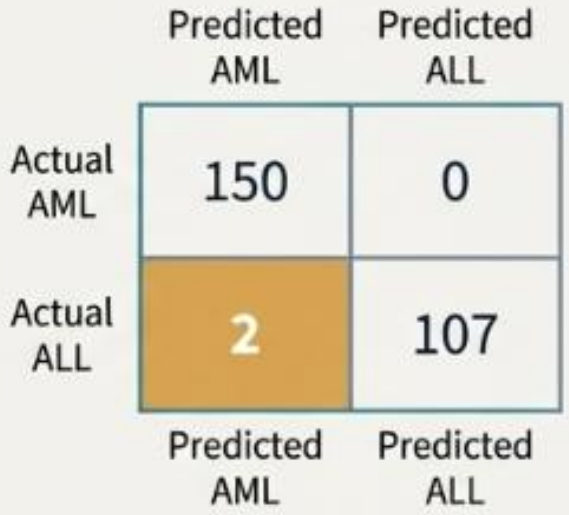
A Deeper Look at Misclassifications

Linear SVM (Lowest Error)



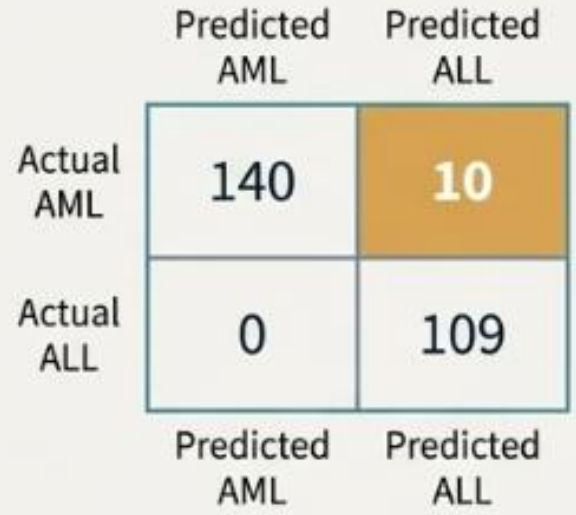
Error Pattern: A single AML case misclassified as ALL. Minimal error, highly specific.

XGBoost (Different Pattern)



Error Pattern: The only model to misclassify ALL as AML. Perfect AML specificity but imperfect ALL sensitivity.

Deep Neural Network (Highest Error)

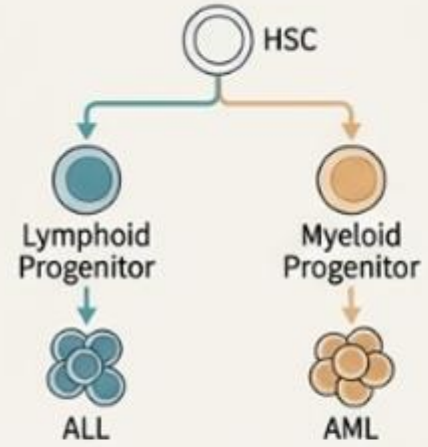


Error Pattern: Amplifies the common AML –> ALL error type, making 10 false positive mistakes.

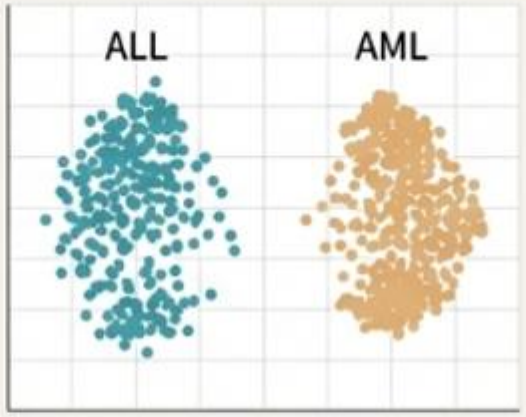
Key Insight: Most models find it easier to mistake an AML sample for ALL than the reverse, suggesting ALL has a more dominant transcriptomic signature. The DNN is particularly susceptible to this bias, learning a less precise decision boundary.

The Biological Truth

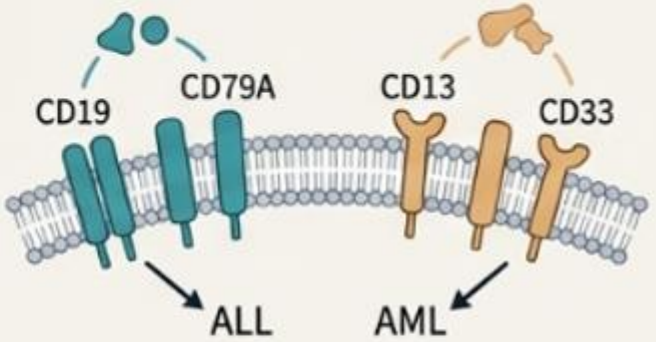
The near-perfect classification is not a statistical artifact; it reflects deep biological reality.



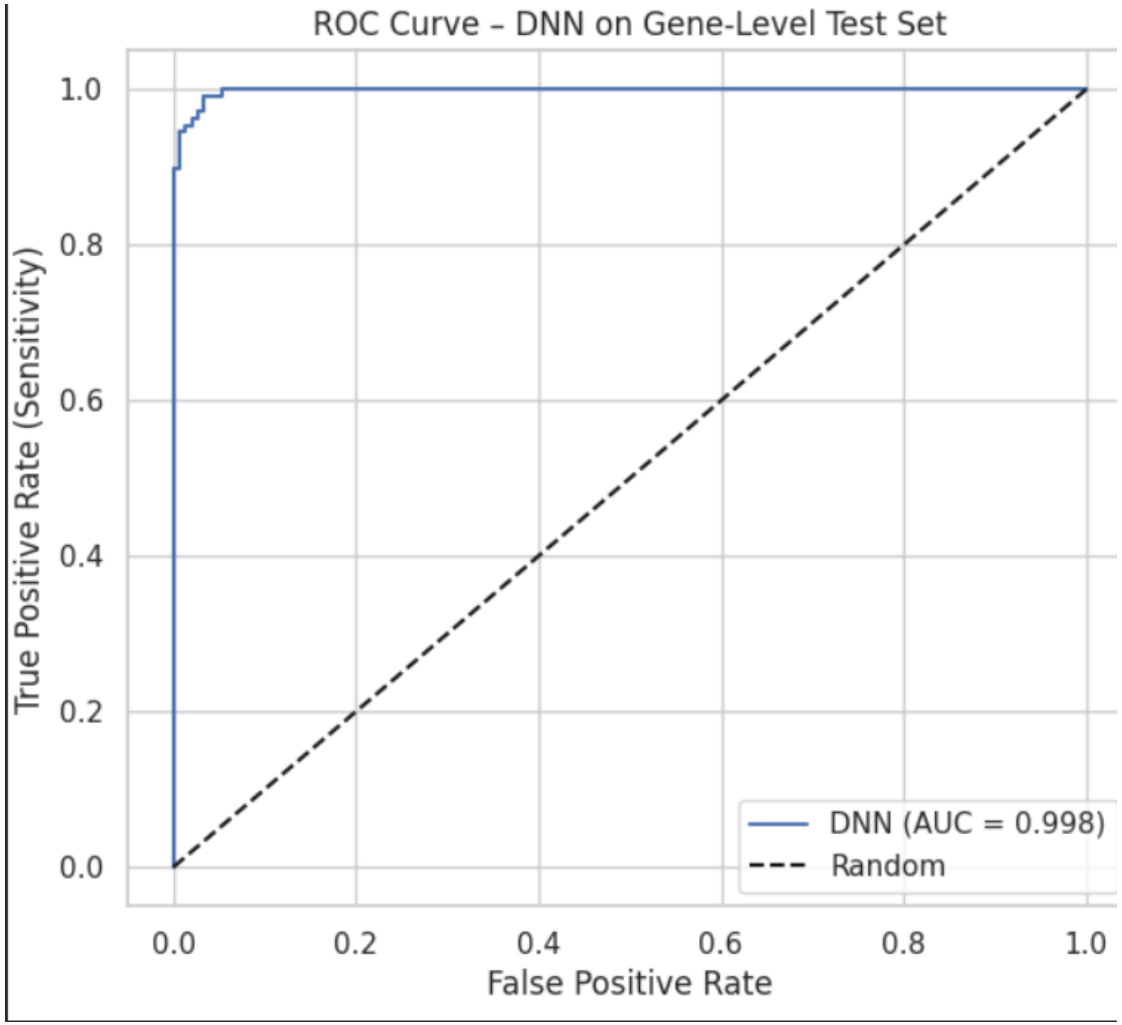
Distinct Cellular Lineages

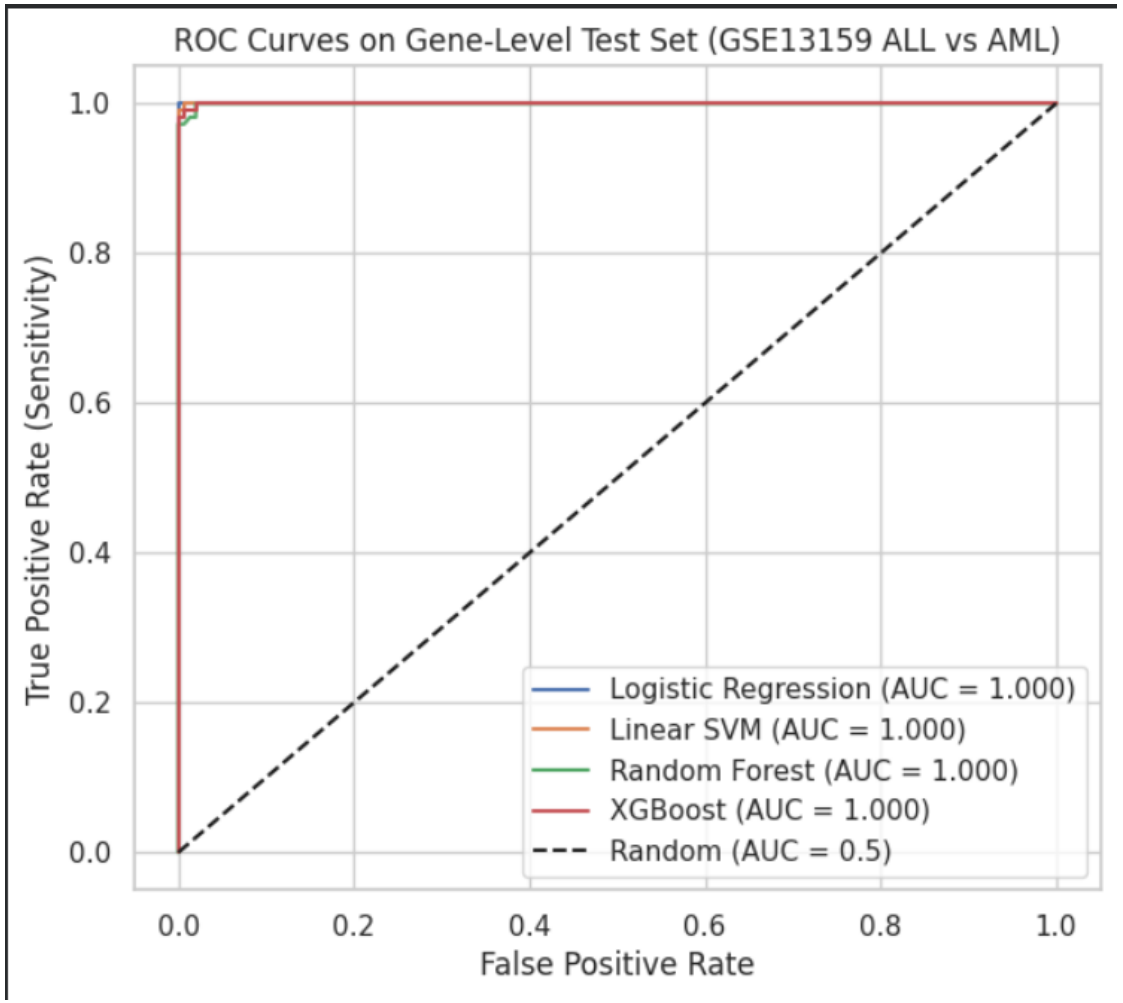


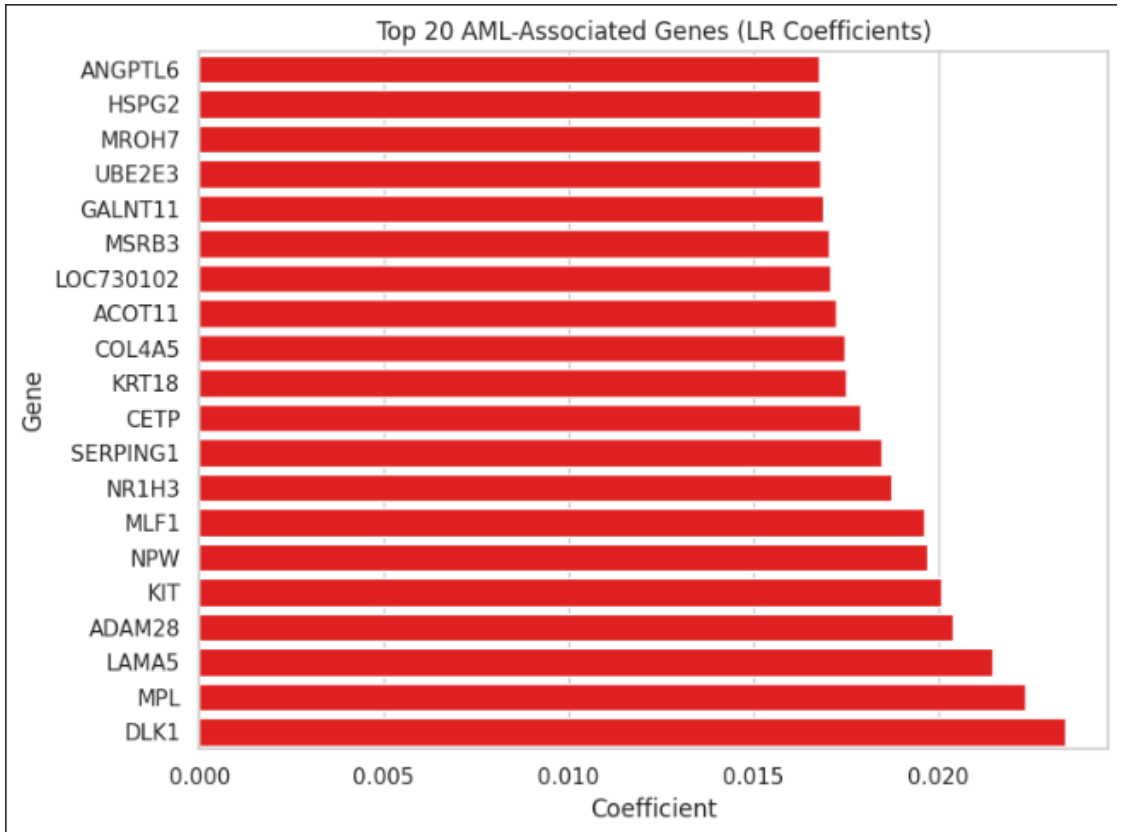
Large Transcriptomic Distance

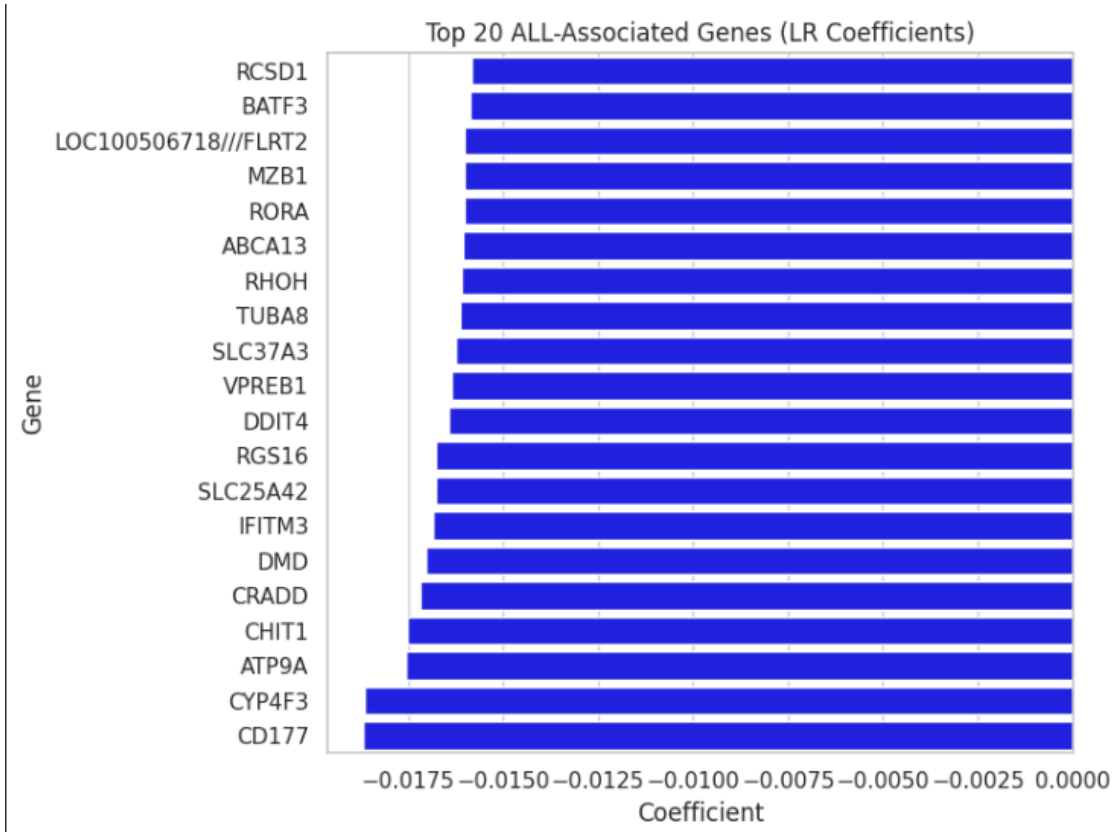


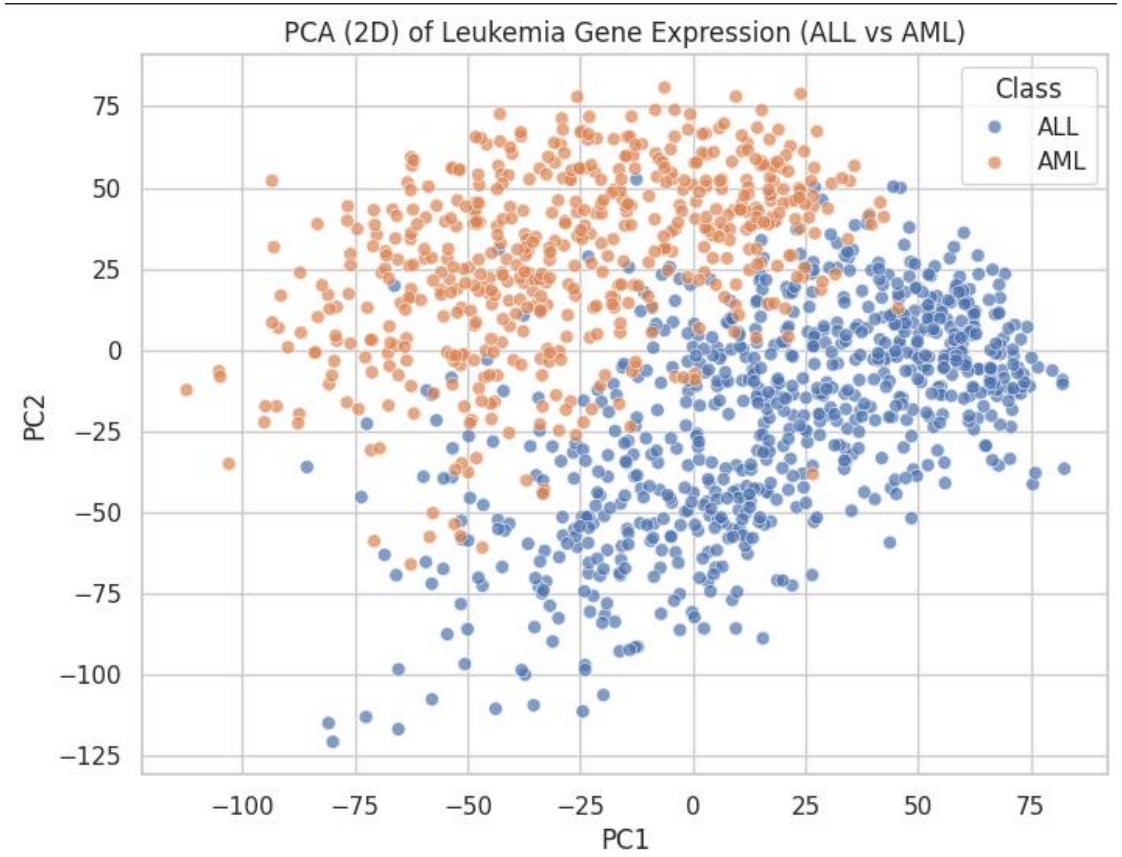
Well-Established Biomarkers

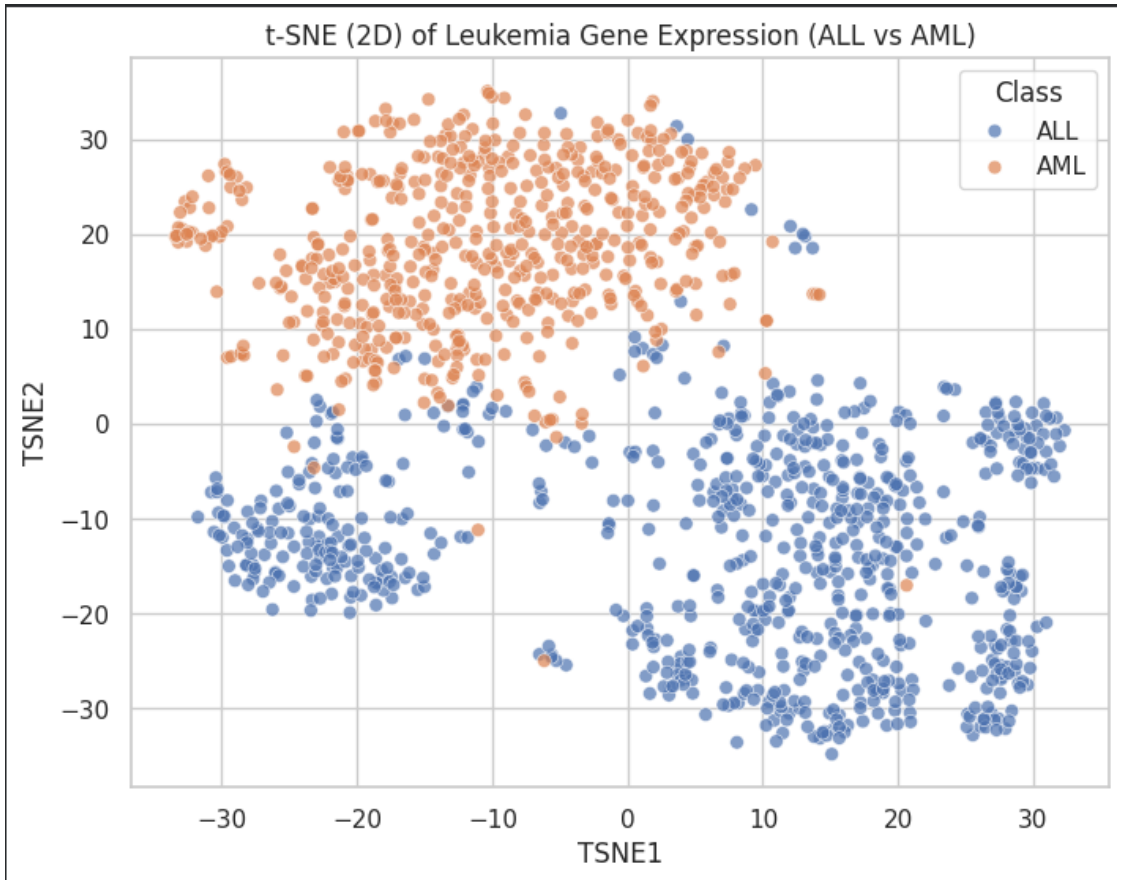

ROC Curve – DNN on Gene-Level Test Set
True Positive Rate (Sensitivity)
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
DNN (AUC = 0.998)
Random


ROC Curves on Gene-Level Test Set (GSE13159 ALL vs AML)
True Positive Rate (Sensitivity)
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 1.000)
Linear SVM (AUC = 1.000)
Random Forest (AUC = 1.000)
XGBoost (AUC = 1.000)
Random (AUC = 0.5)

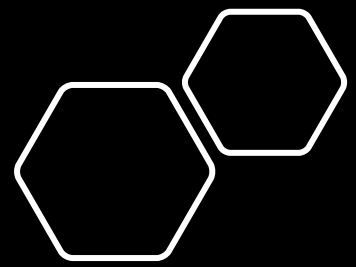

Top 20 AML-Associated Genes (LR Coefficients)
ANGPTL6
HSPG2
MROH7
UBE2E3
GALNT11
MSRB3
LOC730102
ACOT11
COL4A5
KRT18
CETP
SERPING1
NR1H3
MLF1
NPW
KIT
ADAM28
LAMA5
MPL
DLK1
Gene
0.000
0.005
0.010
0.015
0.020
Coefficient


Top 20 ALL-Associated Genes (LR Coefficients)
RCSD1
BATF3
LOC100506718///FLRT2
MZB1
RORA
ABCA13
RHOH
TUBA8
SLC37A3
VPREB1
DDIT4
RGS16
SLC25A42
IFITM3
DMD
CRADD
CHIT1
ATP9A
CYP4F3
CD177
Gene
−0.0175
−0.0150
−0.0125
−0.0100
−0.0075
−0.0050
−0.0025
0.0000
Coefficient


PCA (2D) of Leukemia Gene Expression (ALL vs AML)
Class
ALL
AML
PC1
PC2
75
50
25
0
−25
−50
−75
−100
−125
−100
−75
−50
−25
0
25
50
75


t-SNE (2D) of Leukemia Gene Expression (ALL vs AML)
Class
ALL
AML
TSNE1
TSNE2
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

```
Train shape: (1033, 22189)  Test shape: (259, 22189)
Sample 1
True label: AML
Response : {'prediction': 'AML', 'probability_AML': 0.9320649012916876}
--------------------------------------
Sample 2
True label: ALL
Response : {'prediction': 'ALL', 'probability_AML': 0.00041785078041755593}
--------------------------------------
Sample 3
True label: ALL
Response : {'prediction': 'ALL', 'probability_AML': 0.0017868731551584228}
--------------------------------------
Sample 4
True label: ALL
Response : {'prediction': 'ALL', 'probability_AML': 0.005032610131290868}
Sample 6
True label: AML
Response : {'prediction': 'AML', 'probability_AML': 0.9942616292851311}
--------------------------------------
Sample 7
True label: AML
Response : {'prediction': 'AML', 'probability_AML': 0.966967076128142}
--------------------------------------
Sample 8
True label: AML
Response : {'prediction': 'AML', 'probability_AML': 0.9999982053366313}
--------------------------------------
Sample 9
True label: ALL
Response : {'prediction': 'ALL', 'probability_AML': 0.00018722333274173215}
--------------------------------------
Sample 10
True label: AML
Response : {'prediction': 'AML', 'probability_AML': 0.9999892936651619}
--------------------------------------
```



Future Enhancements:

While highly successful, this blueprint can be further improved with more advanced techniques.

Systematic Feature Selection

Hyperparameter Tuning

Ensemble Methods

Biological Validation