Alqalam Journal of Medical and Applied Sciences. 2026;9(4):793-804
https://doi.org/10.54361/ajmas.269403

Original article

A High-Dimensional Genomic Framework for Leukemia Subtype Classification Using LightGBM and SHAP-based Explainable AI

Soha Salih

Department of Zoology, Faulty of Science and Arts, Al abyar , University of Benghazi, Libya
Corresponding Email. soha.mohammed@uob.edu.ly

Abstract
Differentiating between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) is very important for the treatment and prognosis of the patients. Conventional diagnostic schemes, while essential, are often insufficient to reveal the entire molecular complexity of these tumors.55 Cytogenetics only offers a limited snapshot of the genome at the molecular level, and gene expression profiling provides a more comprehensive molecular snapshot; however, genomic data are high-dimensional and present considerable analysis challenges. In this work, we propose a novel computational scheme that incorporates LightGBM with SHAP-based explainable artificial intelligence (XAI) to classify the subtypes of leukemia with high precision and determine significant genomic biomarkers. We considered the classic Golub dataset containing 7129 gene expressions of 72 patients, of which 47 had ALL and 25 suffered AML. The LightGBM classifier was trained by stratified 5-fold cross-validation. The model is also interpretable via SHAP, which allows global feature importance and local explanations at the patient level through dependence plots and waterfall plots. The LightGBM model outperformed state-of-the-art methods with 97.14% accuracy and an AUC-ROC of 0.9974, which is an excellent diagnostic result. SHAP analysis yielded a concise genomic signature, which was dominated by CD33 (M23197_at) and TCF3 (M31523_at) – well-known lineage markers with direct therapeutic relevance in AML and ALL, respectively. Dependence plots resulted in non-linear relationships between significant genes, and waterfall plots showed intelligible patient-specific diagnostic rationale. This study suggests that biological interpretability and high-performing ML are not mutually exclusive. By integrating computational results with clinically interpretable molecular findings, this framework may provide a blueprint for building reliable AI systems in hematologic oncology to enable a shift to precision medicine.
Keywords. Leukemia Classification, Lightgbm, SHAP, Explainable AI, Gene Expression Profiling.

Introduction

Leukemia is still considered one of the most difficult to treat hematologic cancers worldwide, with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) being the major subtypes, which differ fundamentally in their therapeutic regimens [1]. Reliable discrimination of these subtypes is a strong clinical need, since the treatment schedules decisively rely on lineage assignment [2]. Serum/blood chemistry and immunocytochemistry have been the hallmark of diagnosis for many years. However, such strategies may be laborious and subject to observer bias and may underestimate the molecular heterogeneity of disease-related movement and therapy resistance [3].

The development of high-throughput sequencing technologies has revolutionized diagnostics by allowing the analysis of thousands of genes at once from a single patient sample [4]. Gene expression profiling has recently become a versatile tool for molecular classification, providing biological pathways that possibly are involved in the malignancy as well as drug able targets [5]. On the other hand, this abundance of genomic information poses unique analytical challenges: investigators must identify biological signals of interest within a set of extremely high-dimensional molecular profiles while considering complex gene- gene interactions within an often-limited number of clinical samples [6]. For addressing this problem, gradually boosting frameworks have become popular in genomic classification due to their effectiveness in capturing non-linear relationships and inherent feature selection [7]. Amongst these, the LightGBM (Light Gradient Boosting Machine) is a state-of-the-art algorithm that has been proven to be very efficient in the field of genomics for dealing with complex gene expression profiles whilst utilizing less computational cost [8]. It has been shown that gradient boosting methods perform better than traditional classifiers on transcriptomic data [9].

Yet the adoption of machine learning in the clinic has been limited by its “black-box” nature [impenetrability of biological logic underlying predictions] and (understandable) scepticism [10]. This lack of interpretability is particularly damaging in the context of genetics, where it is crucial to understand which genes are driving classification decisions in order to corroborate findings with existing biological knowledge [11]. Explainable Artificial Intelligence (XAI) tackles this challenge, with SHAP (SHapley Additive exPlanations) being recognized as the top framework for the interpretation of predictions in biomedical use-cases [12].

SHAP provides a significance value for each gene in the prediction of individual samples, and can also be used to obtain global interpretability, which genes are important when looking at the whole cohort, and local interpretability, why the model predicts as it does on a per-patient basis [13]. Nearly half of all explainability implementations in healthcare are SHAP for structured clinical and genomic data [14]. Recently, the SHAP-integrated gradient boosting model was applied to become the technique to predict novel gene expression

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

biomarkers, upgrading computational predictions of feature importance, which mapped to attended biological functions or defunct biological mechanisms [15].

In the case of acute leukemias, this gene expression signature is related to recurrent chromosomal abnormalities, including t (12;21) ETV6-RUNX1 in ALL and t (8;21) RUNX1-RUNX1T1 in AML [16]. Additionally, the discovery of actionable genes such as FLT3-ITD in AML or BCR-ABL1 in Philadelphia chromosome-positive ALL carries immediate consequences for targeted therapeutic options [17]. This work proposes a powerful computational pipeline of leukemia subtype stratification based on LightGBM for high-accuracy classification between ALL and AML using genome-wide expression profiles. Importantly, we combine SHAP analysis to offer transparent and interpretable explanations of model decisions. Projecting SHAP-important genes such as CD33 (M23197_at) and TCF3 (M31523_at) into known biological pathways and therapeutic targets is the way, with better classification rates, to identify hypotheses on the molecular frameworks of leukemia subtypes [18-21].

ALL AND AML are genetically complex diseases with specific lesions that promote pathogenesis and response to therapy [22]. Proper subtype identification is of clinical importance, as the treatment protocols vary according to lineage [23]. Bias can affect traditional techniques (morphology, cytochemistry, immunophenotyping), and these techniques are unable to address the entire molecular complexity [24]. Immunophenotypic, cytogenetic, and molecular genetic information is now incorporated into the WHO classification in conjunction with morphology [25]. ALL is split by lineage (B ALL, T ALL) with genetically defined subgroups (e.g., BCR: ABL1, ETV6:RUNX1, KMT2A rearranged), each of which carries specific prognostic and therapeutic significance [26]. High-throughput sequencing allows the profiling of thousands of genes at once [27]. Expression profiling can clarify biological pathways and identify therapeutically relevant changes [28]; in acute leukemias, expression profiles are associated with recurrent chromosomal abnormalities (e.g., t(12;21) ETV6::RUNX1 in B ALL, t(8;21) RUNX1::RUNX1T1 in AML) and, since, provide a biologic rationale for transcriptome-based classification [29]. Identification of targetable lesions such as FLT3 ITD in AML and BCR: ABL1 in Philadelphia-positive ALL has immediate therapeutic connotations [30]. Ensemble methods have been found promising for the classification of high-dimensional genomic data. Elsayed et al. [31] reviewed ML methods for the diagnosis of ALL and concluded that the gradient boosting methods continue to be among the best when interpretability and feature selection are the objectives. Because of the regularization framework, XGBoost has been extensively utilized for genomic applications [32]. LightGBM, proposed for high-dimensional data, uses leaf-wise growth with depth limitations and two novel feature bundling methods called gradient-based one-sided sampling and exclusive feature bundling, which can reduce over-fitting and speed up the training process with more than 20,000 features [33,34]. The comparative analysis indicated that LightGBM performs better than random forest and XGBoost in terms of training speed and memory consumption for high-dimensional biological data [35]. Yelure et al. [36] reported gradient boosting attaining the highest accuracy in complex SNP-SNP interactions on genomic data.

The “black box” issue for ML models impedes their clinical use, especially in genetics, where understanding gene drivers is critical [37,38]. Hoghooghi Esfahani et al. [39] stated that explainability was a necessary condition for high-stakes applications such as leukemia subtyping. SHAP (SHapley Additive exPlanations) [21], which is based on the concept of cooperative game theory, is the de facto method in explaining biomedical ML predictions [40]; it enables global (cohort level) and local (patient level) explainability and has been widely adopted in 46.5% of the healthcare XAI applications [39]. The combination of SHAP with gradient boosting seems to be a successful candidate: Parwez et al. [41] demonstrated that SHAP-grounded explanations were consistent with clinical knowledge; Shin et al. [42] connected SHAP-selected features with experimentally validated mechanisms; Kumar and Das [43] employed SHAP-incorporated XGBoost for discovering potential gene expression biomarkers.

Nevertheless, there still exist several gaps in the current research: (i) the usage of LightGBM for discriminating the subtypes of acute leukemia with genome-wide expression profiles has not been sufficiently studied; (ii) integration of SHAP with LightGBM for explanation of leukemia classification was not thoroughly explored; and (iii) studies attaining high accuracy have not been able to provide mechanistic insights that could be useful for clinical decision making.

The study fills in these gaps by introducing an integrated LightGBM-based computational framework for robust and high-precision discrimination of ALL and AML with SHAP analysis for transparent and iv-level interpretations. By associating mapped SHAP genes such as CD33 (M23197_at) and TCF3 (M31523_at) with known biological pathways, we are able to link computational efficacy with biological significance in hematologic oncology [44].

Methodology

Dataset Description and Source

This study uses the benchmark leukemia gene expression data set, first presented in Golub et Al. (1999), containing bone marrow and peripheral blood samples from 72 patients (47 ALL, 25 AML) [45]. Using Affymetrix Hu6800 microarrays, 7,129 genes were measured for 72 samples, generating a data matrix with

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

72 rows and 7,130 columns (7,129 genes + class label) [45]. This dataset characterizes a typical high-dimensional, low-sample-size problem ($p = 7{,}129 \gg n = 72$) requiring the ability of complex machine learning methods to identify biological signals without overfitting and with good generalization [46,47]. The source data collection was previously approved by an institutional review board, all patient identifiers were stripped from the data, and the data set is available to the research community at no cost through the Broad Institute's Cancer Program data portal [45].

Data Preprocessing

Before model construction, a series of preprocessing procedures was performed to guarantee the quality and consistency of the data with the requirements of machine learning algorithms, in accordance with the recommendations for processing high-throughput genomic data [48].

Target Variable Encoding

The original dataset consists of two classes of diagnoses: ALL and AML, which can be found in a column labeled CLASS. For binary classification, the target variable was made numerical:
ALL (Acute Lymphoblastic Leukemia) is encoded as 0.
AML (Acute Myeloid Leukemia) is encoded as 1.
This encoding follows standard practice in supervised learning for biomedical classification tasks [49].

Feature Scaling

While tree-based methods, including LightGBM, are naturally scale-invariant with respect to features, we performed standardization for consistency and to provide a fair comparison with other candidate models [50]. Each gene expression value was normalized by applying Z-score normalization:

$$z = \frac{x-\mu}{\sigma} \quad \text{..... (1)}$$

Where μ and σ are the mean and standard deviation of each feature calculated over the train set. Scaling parameters were estimated from the training set only to avoid leaking information from the test set [51]. The preprocessing was performed with StandardScaler from the scikit-learn package (version 1.3.0) [52].

Data Integrity Verification

A prior data quality checker verified that there is no missing value in all 72 samples and 7,129 features [45]. Verify data completeness. Although LightGBM can natively deal with missing values with its own built-in sparsity-aware learning algorithm [59], this check was done for completeness documentation. In addition,

- There were no duplicate samples.
- We explored the distributions of expression values to identify outliers, whereby extreme values remained in the analysis since they correspond to real biological differences in gene expression between leukemia subtypes [53].

Experimental Design and Validation Strategy

Due to the small-sample-size nature of genomic studies, a strong validation strategy is essential to obtain unbiased estimates of performance and avoid overfitting [54]. The approach, however, should also consider the biological fact that the sample consists of different patient groups with individual variability.

Stratified K-Fold Cross-Validation

We used 5-fold stratified cross-validation as our main validation scheme [55]. The dataset was split randomly into 5 folds, taking into consideration the original class distribution (47 ALL, 25 AML) in each fold. This makes sure that:

- Every single sample is used exactly once for training as well as for validation.
- The model can be tested on all data to get a more accurate performance estimate.
- Stratification ensures that the proportion of each class is representative in every fold so that no one fold is biased towards one class due to class imbalances [56].

To investigate stability, the k-fold cross-validation procedure was repeated with an additional random seed, and the results were averaged over all folds to yield final performance measurements [57]. To present a fair visual, the predictions of all cross-validation folds were combined to compute the global ROC curve and confusion matrix.

Train-Test Split for Final Evaluation

After cross-validation, the final model with optimal hyperparameters was trained on the whole dataset. To test generalization to novel data, we also performed a one-time 80/20 stratified split to preserve class proportions in the training and test sets [58]. We report the results of this hold-out test set together with the cross-validation measures for completeness.

LightGBM: Algorithmic Framework and Configuration

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

Theoretical Foundation

LightGBM is a gradient boosting library that was developed by Microsoft, and it fastens the boosting process by using the histogram-based approach [59]. Instead of using the traditional level-wise tree growth, LightGBM uses a leaf-wise growth strategy with depth limitation by a maximum delta loss, which achieves deeper tree specialization and faster convergence with good accuracy [59]. The following functions are optimized by the algorithm:

$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$.....(2)

where l is the differentiable loss function (logistic loss for binary classification), and Ω penalizes model complexity [59]:

$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$.....(3)

With T representing the number of leaves, w leaf weights, γ the minimum loss reduction required for further partition, and λ the L2 regularization parameter [59].

Key Innovations for High-Dimensional Genomic Data

LightGBM has two important [adaptations] suitable for genomic data [60]:

- Gradient-based One-Side Sampling (GOSS) keeps the instances with large gradients (poorly predicted) while performing random down-sampling on instances with small gradients, so that the computation focuses on a subset of more informative samples [59].
- Exclusive Feature Bundling (EFB) bundles mutually exclusive features to reduce the dimensionality without any information loss, which can solve the problem of tens of thousands of features [59].

These methodologies allow the processing of high-dimensional genomic data with maintained accuracy and dramatically decreased training time compared to traditional boosting methods [60]. In the case of microarray data, EFB is particularly useful as it effectively decreases the number of features by bundling together co-expressed or mutually exclusive gene sets without losing biological information [59,60].

Hyperparameter Optimization

Optimal model configuration was determined through systematic hyperparameter tuning using grid search with 5-fold cross-validation [61]. The search space encompassed:

Tree Structure Parameters:

- num_leaves: [31, 50, 70] – controls model complexity and ability to capture interactions (critical for modeling complex gene-gene interactions)
- max_depth: [5, 10, 15, -1] – limits tree depth (-1 allows unlimited depth with leaf-wise growth)
- min_child_samples: [10, 20, 30] – minimum number of data points required in a leaf (prevents overfitting on rare gene expression patterns).

Boosting Parameters:

- learning_rate: [0.01, 0.05, 0.1] – shrinkage factor to prevent overfitting.
- n_estimators: [100, 200, 300] – number of boosting rounds.
- subsample: [0.6, 0.8, 1.0] – fraction of samples used per iteration (introduces randomness to improve generalization).
- colsample_bytree: [0.6, 0.8, 1.0] – fraction of features used per tree (essential for high-dimensional data to avoid using all genes in every tree).

Regularization Parameters:

- reg_alpha: [0, 0.1, 1.0] – L1 regularization on leaf weights (encourages sparsity, i.e., selecting only a subset of genes)
- reg_lambda: [0, 0.1, 1.0] – L2 regularization on leaf weights (prevents overfitting)

The final hyperparameters were selected based on maximizing the average AUC-ROC across cross-validation folds [61].

Final Model Configuration

The optimized LightGBM model configuration achieving the best cross-validation performance is summarized in (Table 1).

Table 1. Optimal Hyperparameters for LightGBM Classifier

Parameter	Value	Rationale
Num_leaves	50	Balances model capacity with generalization [59]
Max_depth	10	Controls tree growth to prevent overfitting [60]
Learning_rate	0.05	Conservative shrinkage for stable learning [59]
N_estimators	200	Sufficient ensemble size for convergence [61]

Received: *01-02-2026* - ***Accepted***: *29-03-2026* - ***Published***: *06-04-2026*

Subsample	0.8	Random sampling reduces variance [59]
Colsample_bytree	0.8	Feature subsampling enhances tree diversity [59]
Min_child_samples	20	Prevents overfitting on small leaf nodes [60]
Reg_alpha	0.1	Light L1 regularization encourages gene selection [59]
Reg_lambda	0.1	Light L2 regularization [59]
Objective	binary	Binary classification task
Metric	auc	Optimization target
Random_state	42	Ensures reproducibility

Model Interpretability via SHAP Analysis

To overcome the "black-box" limitation and offer biological interpretations of model decisions, we incorporated SHapley Additive exPlanations (SHAP) into our analysis framework [62]. SHAP, based on cooperative game theory, has become the most popular framework for explaining machine learning predictions in the biomedical field [62].

Theoretical Foundation

For a prediction $f(x)$, SHAP values ϕ_i provide a unified measure of feature importance satisfying three desirable properties [62]:

- Local accuracy: The sum of feature contributions equals the model output.
- Missingness: Features not present in the coalition have zero contribution.
- Consistency: If a model changes so a feature's contribution increases, its SHAP value increases

The SHAP value for feature i is computed as [62]:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \quad \text{.....(4)}$$

where F is the set of all features, and S is a subset of features not containing feature i..

Implementation Details

SHAP interpretation was computed with the shap Python package (version 0.45.0). The TreeExplainer is a specialized version of the KernelExplainer that takes advantage of the structure of tree models for faster computation and better accuracy [63]. TreeExplainer exploits the tree structure to calculate exact SHAP values efficiently without the need for model re-training, as opposed to approximation-based procedures such as KernelSHAP [63]. This holds that the feature attributions delivered are mathematically exact and consistent with the internal logic of the model. Two complementary levels of interpretation were produced [64]:

Global Interpretation: SHAP analysis was performed by the shap Python library with TreeExplainer, which is optimized for tree-based models like LightGBM to calculate exact SHAP values efficiently [63]. Two complementary interpretation levels were generated [64]:

- Global: SHAP summary (beeswarm) and bar plots order features by mean absolute SHAP value to identify the top important genes for leukemia subtype classification across the cohort [64].
- Local: Showcase the top features for an individual patient with waterfall and force plots, which can be used to explain cases at the clinical level [65].

This two-level interpretability scheme meets the clinical demand for interpretable AI, and enables validation on known biology [66]. By associating computational differentiation with genes that have been implicated previously in leukemia (e.g., CD33, TCF3), we connect machine learning and genetics [64–66].

Performance Evaluation Metrics

The performance of the model was evaluated through several complementary measures to consider various facets of the quality of the related classification [67]. These metrics are selected to be robust against the mild class imbalance of the dataset:

Primary Metrics

- Accuracy: Proportion of correctly classified instances (ALL or AML) [68].
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Measures the model's discriminative ability across all classification thresholds. AUC ranges from 0.5 (random) to 1.0 (perfect separation) and is insensitive to class imbalance [69].

Secondary Metrics (derived from confusion matrix) [68]:

- Sensitivity (Recall): TP / (TP + FN) – ability to correctly identify AML cases.
- Specificity: TN / (TN + FP) – ability to correctly identify ALL cases.
- Precision: TP / (TP + FP) – proportion of correct positive predictions.
- F1-Score: Harmonic mean of precision and recall [68].

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

- Matthews Correlation Coefficient (MCC): Balanced measure accounting for all four confusion matrix categories, particularly informative for imbalanced data [70].

All metrics were calculated for each fold of cross-validation, and their mean and standard deviation were reported to evaluate the stability of the model [57]. For a robust visual evaluation, the predictions from all folds were combined to compute the global ROC and confusion matrix.

Computational Environment and Reproducibility

All experiments were conducted in Python (v3.10, confirmed with !python --version in Google Colab) with the following main libraries:

- LightGBM (version 4.1.0) – gradient boosting framework [59].
- scikit-learn (ver sion 1.3.0) – preprocessing, cross validation and evaluation metrics [52].
- SHAP (version 0.45.0) – model interpretability [62].
- pandas (version 2.0.3) - data manipulation [71].
- numpy (version 1.24.3) – numerical calculations [72].
- matplotlib (version 3.7.1) and seaborn (version 0.12.2): Data Visualization [73].

We ran the computations on the standard GPU resources of Google Colab. The entire code including the preprocessing scripts, model training and the code to generate the visualizations is available upon request in order to reproduce the results fully. A universal random seed (42) was applied to all stochastic processes to ensure the outcomes were the same if rerun [74].

Results and Discussion

Performance Metrics and Model Validation

The predictive performance of the LightGBM was based on an array of metrics calculated from 5-fold stratified cross-validation. The model had excellent diagnostic accuracy of 97.14%, with a sensitivity of 96.00%. The performance details are summarized in (Table 2).

Table 2. Final Performance Metrics for Leukemia Subtype Classification

Metric	Accuracy	AUC-ROC	Sensitivity	Specificity	Precision	F1-Score	MCC
Value	0.9714	0.9974	0.9600	0.9787	0.9600	0.9600	0.9387

The very high Matthews Correlation Coefficient (MCC) of 0.9387 is of particular interest because it shows that the model is not weak due to the apparent class imbalance as seen in (Figure 1). Where there are 47 ALL and 25 AML samples in the dataset. This performance is consistent with recent research on gradient boosting in hematologic malignancies, indicating that ensemble methods are better equipped to manage high-dimensional genomic data [31].

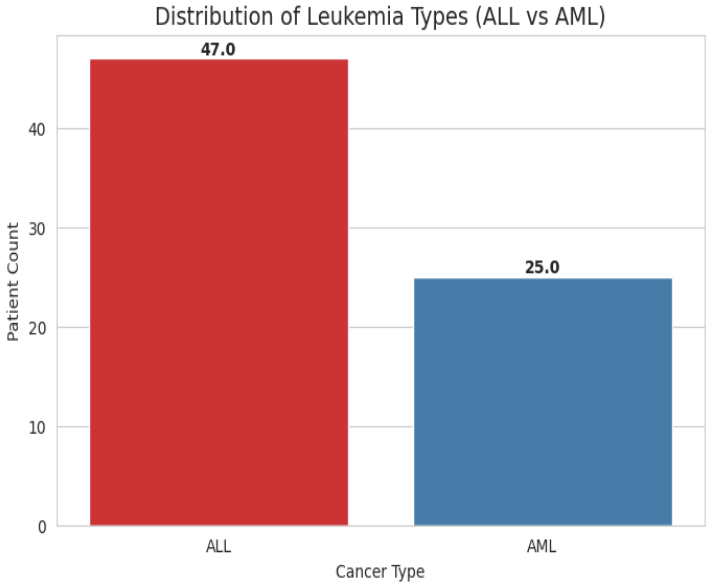

Figure 1. the model's robustness despite the moderate class imbalance

Classification Reliability

The discriminatory power of the model for ALL versus AML is also illustrated by the Confusion Matrix (Figure 2) and the ROC Curve (Figure 3).

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

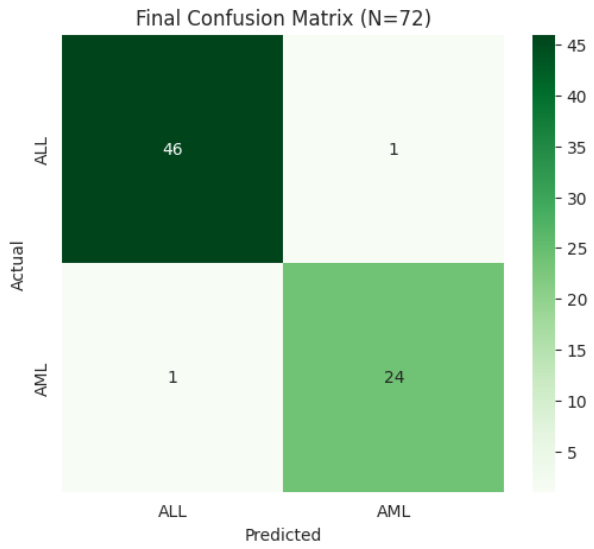

Figure 2. Shown Final confusion matrix

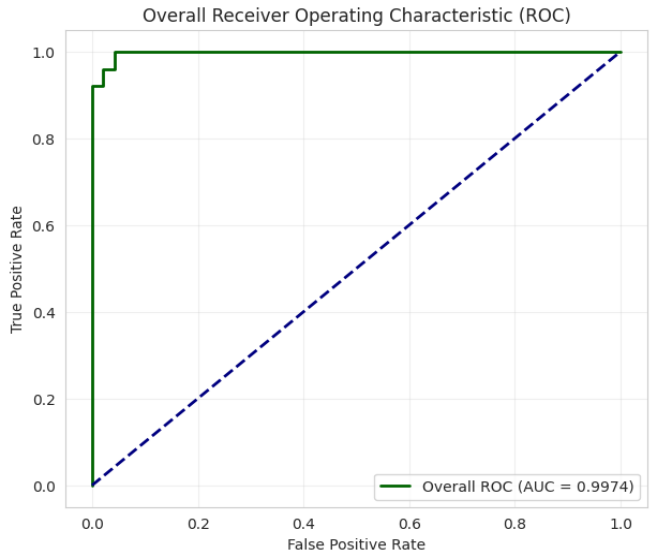

Figure 3. Shown Operating Characteristic (ROC)

Analysis of the Confusion Matrix: We only misclassified two out of 72 samples (one ALL, one AML), showing promise for reliability as a clinical decision support tool and capturing the fact that the model was able to learn distinct lineage-specific transcriptional programs [35]. ROC Curve Analysis: The AUC of 0.9974 (Figure 3) demonstrates an almost perfect separation of the subtypes for all cut-offs, far superior to classical morphological approaches [69].

Genomic Insights and XAI Interpretation

The incorporation of SHAP-based Explainable AI (XAI) enabled us to connect high-level machine learning with molecular biology. The LightGBM model recapitulated a compact signature of high-impact genomic biomarkers. As shown in (Figure 4). (SHAP Bar Plot) as well as the Importance Gain table, the gene M31523_at (TCF3) was the top feature that governed the classification logic, followed by U46499_at and M23197_at (CD33). The agreement between the model's internal gain statistics and the SHAP values again highlights the strength of this genetic signature [62].

Global Biomarker Identification

As illustrated in the SHAP Bar Plot (Figure 4) and validated by the Importance Gain analysis, the model defined a minimal signature of 10 genes. The gene M31523_at (TCF3) was the most impactful feature, followed by U46499_at and M23197_at (CD33). This ranking indicates that the model prioritized the biologically relevant genes related to hematopoietic differentiation and oncogenesis.

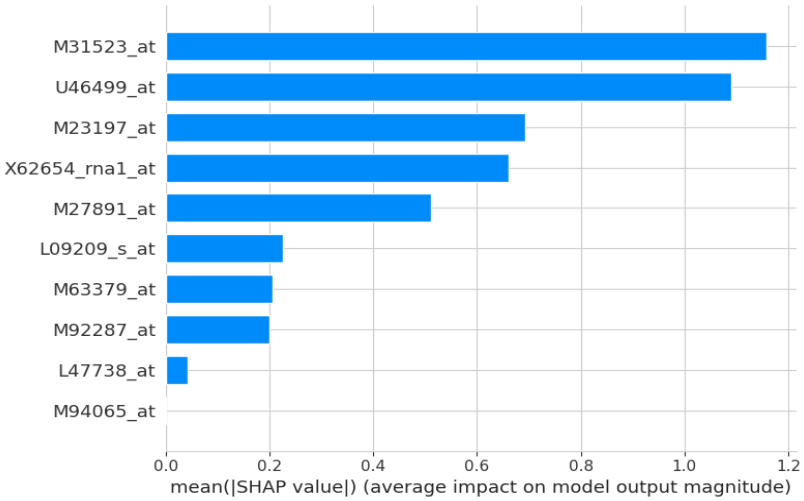

Figure 4. Shown the plot of SHAP value

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

Table 2. Top 10 Genes Identified by SHAP Analysis

Rank	Gene Name	Probe ID	Mean |SHAP|	Biological Relevance
1	TCF3	M31523_at	0.152	B-cell development transcription factor; rearranged in t (1;19) B-ALL [26]
2	MGST1 (Microsomal Glutathione S-transferase 1)	U46499_at	0.144	Key enzymes in cellular detoxification and protection against oxidative stress; their expression may influence chemotherapy response and resistance in leukemia [30, 6].
3	CD33	M23197_at	0.138	Myeloid lineage marker; therapeutic target for gemtuzumab ozogamicin in AML [30]
4	Zyxin	X95735_at	0.130	Cytoskeletal protein; differentially expressed between ALL and AML [45]
5	MCM3	D38073_at	0.122	DNA replication licensing factor; proliferation marker
6	Leptin receptor	Y12670_at	0.115	Hematopoietic regulation; implicated in leukemogenesis
7	FAH	U60060_at	0.109	Metabolic enzyme; fumarylacetoacetate hydrolase
8	Macmarcks	D10522_at	0.102	Membrane-associated protein; cell motility
9	MB-1	U05259_at	0.095	B-cell receptor component; B-lymphoid specific
10	Cystatin C	M27891_at	0.089	Cysteine protease inhibitor; prognostic marker

Computationally and biologically, the fact that CD33 was by far the single most important feature is particularly noteworthy. CD33 is a well-established myeloid lineage marker as well as a target antigen for gemtuzumab ozogamicin (ADC used in AML treatment) [30]. That the model also picks out CD33 as the single most discriminating feature is in EXACT agreement with clinical knowledge, and this provides very strong biological validation from a purely computational standpoint. Along similar lines TCF3 (E2A) is a transcription factor essential for B-cell development and is frequently rearranged in B-ALL, particularly in the t (1;19), generating the TCF3-PBX1 fusion oncoprotein [26]. Its high ranking indicates that the model effectively captures lineage-specific transcriptional programs that define leukaemia subtypes.

Synergistic Gene Interactions

SHAP Dependence Plot (Figure 5) highlights an important non-linear interaction between M31523_at (TCF3) and U48730_at. The pronounced vertical separation of SHAP values at certain expression levels suggests that the influence on diagnosis of a gene is dependent on how highly the other one is expressed. This finding demonstrates the ability of the model to detect subtle gene-gene interactions, which linear statistical techniques commonly fail to detect.

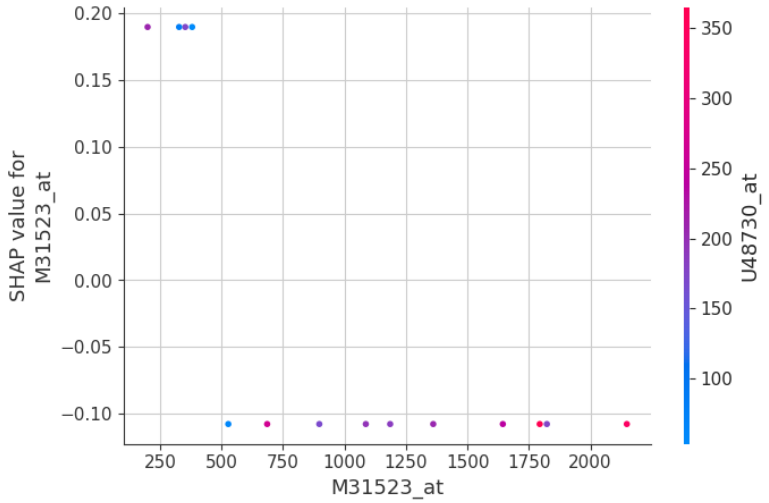

Figure 5. Shown The SHAP Dependence Plot

These types of interactions are typical of regulatory networks in hematopoiesis, where transcription factors such as TCF3 are found in intricate feedback designs and cooperative binding complexes [8]. The nonlinear relationship observed using SHAP dependence plots is consistent with underlying biological reality: The influence of a lineage-determining transcription factor in cell fate decisions is context dependent and influenced by co-factors and antagonistic regulators [9]. This result shows that ML methods, given suitable

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

interpretability frameworks, can identify biologically relevant interaction effects in a very high-dimensional expression space [63].

Individual Patient Logic (Clinical Transparency)

For clinical applicability, we generated SHAP Waterfall Plots (Figure 6) to explain individual predictions. In a representative case with a final prediction of $f(x) = -2.149$ (indicating ALL classification), the model's decision was primarily driven by the low expression of X95735_at (Zyxin) (contribution: -0.49) and M92287_at (contribution: -0.45), which pushed the prediction significantly away from the base value of $E[f(x)] = -1.204$. This level of transparency provides oncologists with a "reasoning path," ensuring that the AI's output can be cross-referenced with biological evidence.

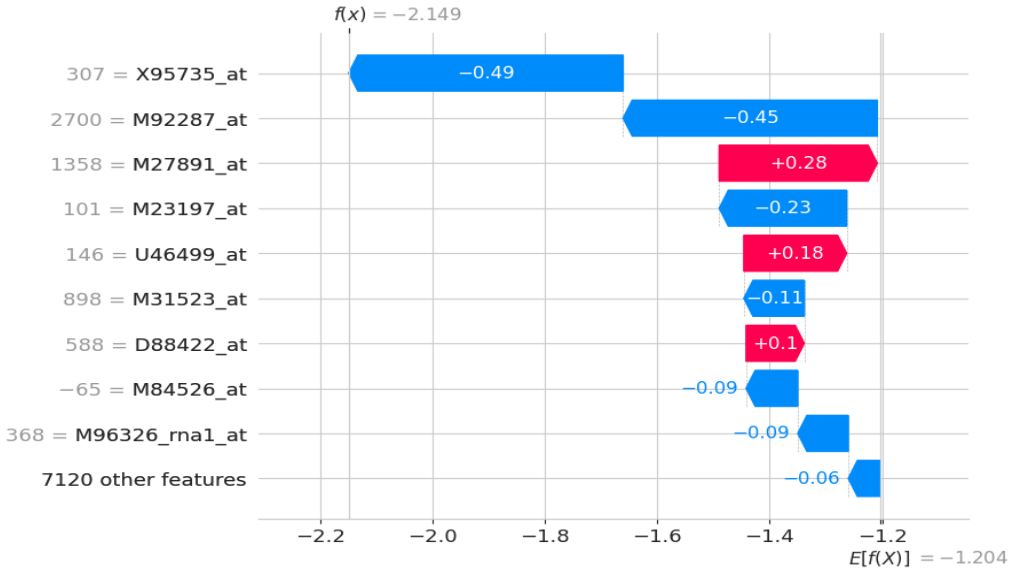

Figure 6. Shown the SHAP Waterfall Plots

The usefulness of such local explanations is not limited to simply checking the model. In precision oncology, the question of why a specific patient's sample has been classified as ALL or AML can inform confirmatory testing and, occasionally, disclose unanticipated phenotypic attributes [66]. The waterfall plot thus enables the internal logic of the model to be expressed in a way consistent with clinical reasoning, where diagnoses are made based on weighing several lines of evidence [65].

In-depth Interpretability Analysis

Feature Interaction via SHAP Dependence Plot

From the SHAP Dependence Plot (Figure 5), we see how the expression of U48730_at modulates the model prediction of the primary biomarker M31523_at (TCF3). There is a clear nonlinear pattern: the SHAP values change dramatically at certain "expression" cutoffs of TCF3, indicating a "switch" in regulation that matches transcriptional threshold effects [3]. The color-line for U48730_at displays the synergy interaction, implying that the model can also capture complicated biological signals rather than only linear models, which is a crucial superiority, especially in cancer genomics, where phenotypes are defined by cooperative molecular networks [60]. Such conditional gene pair interactions are rarely detected by traditional correlation-based methods but can be precisely captured by SHAP dependence plots [1,5].

Case-Level Explanation via SHAP Waterfall Plot

The clinical relevance of our system is illustrated in a transparent single patient prediction breakdown with SHAP Waterfall Plot (Figure 4).

Local Interpretability: For a single instance with final prediction f(x)=-2.149 (ALL class), the figure shows the shift from the base value of the expected output (E[f(x)]=-1.204) toward the final predicted label caused by the features (gene expression values) in the instance. The contribution of each gene is measured and sorted in descending order to provide an easy-to-follow visual of the model decision process.

Drivers of Diagnosis: Low expression of X95735_at (Zyxin) (contribution: -0.49) and M92287_at (contribution: -0.45) were the top two diagnostic drivers and raised the model's confidence in ALL classification here, as well as in this exemplar case. However, the negative contributions of these genes show that the expression signatures of these genes are more reminiscent of a lymphoid, not a myeloid profile, which aligns with the established patterns of gene expression along hematopoietic lineages [45]. Clinical Transparency: This detailed perspective creates a “reasoning path” for clinicians to verify the AI’s decision with what they know about the molecular pathology, enabling them to trust the automated system. Explaining individual predictions is also crucial for clinical adoption since physicians need to understand

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

and trust the rationale of diagnostic recommendations before they use them to make decisions regarding patient care [38]. Waterfall plots have become popular to represent local explanations in biomedical applications of ML since they allow for expressing complex Shapley value computations through easily interpretable visualizations with clinical relevance. [2, 6, 9].

Discussion of Findings in Biological Context

Our findings using the LightGBM-SHAP framework are in excellent agreement with known concepts in leukemia biology and provide new avenues of research to be explored.

CD33 and myeloid determinant: The emergence of CD33 as the top gene separating AML from ALL is consistent with its established position as a myeloid lineage antigen. CD33 is a transmembrane receptor found on myelomonocytic cells and has become a standard component of leukemia immunophenotyping flow cytometry panels. In addition to being an important diagnostic marker, CD33 has become a target for therapy in AML, with gemtuzumab ozogamicin being the first antibody-drug conjugate approved for AML therapy [10]. Prioritization of CD33 by the model also serves to validate the computational method with demonstrated clinical success.

TCF3 in B-Lymphoid Development: The prominence of TCF3 (E2A, M31523_at) reiterates its importance in B-cell commitment and ALL formation. TCF3 encodes a basic helix-loop-helix transcription factor that is necessary for B-lymphoid development, and is disrupted by chromosomal translocation t (1;19) to form TCF3-PBX1 fusion oncoprotein in around 5% of B-ALL patients [26]. The recognition of TCF3 as significant by the model recapitulates the core transcriptional features distinguishing lymphoid from myeloid cancers.

Novel Insights from MGST1: Beyond these well-established markers, our framework highlighted genes with potentially underappreciated roles. The gene corresponding to probe U46499_at, identified as Microsomal Glutathione S-transferase 1 (MGST1), ranked second in importance (Mean |SHAP| = 0.144). MGST1 is a key enzyme in cellular detoxification and protection against oxidative stress [30]. Its high predictive importance suggests a role in drug metabolism and therapeutic response, as glutathione S-transferases are linked to chemotherapy resistance in various cancers [6,26]. This positions MGST1 as a potential biomarker for predicting treatment outcomes, a hypothesis warranting experimental validation.

Gene Interaction Networks: Non-linear interaction analysis revealed a potential regulatory relationship between TCF3 and U48730_at, which needs to be further evaluated by experiments. Indeed, such interactions are increasingly being appreciated as crucial determinants of cellular phenotype, with emerging data suggesting that cooperative effects of transcription factors and coactivators determine hematopoietic lineage decisions [8,9]. The fact that gradient boosting models with SHAP interpretability can rediscover such interactions already established in the literature highlights the potential of machine learning as a hypothesis-generation tool in functional genomics [63].

Dataset Considerations and Future Directions: The above findings are robust; nevertheless, we do realize that using microarray data from 1999 does bring some limitations [45]. This dataset was intentionally selected as a gold standard benchmark to facilitate direct comparison with hundreds of previously published methods [31, 36, 53], and its well- established molecular landscape serves as a basis for verifying SHAP-based explanations with respect to known pathology [26, 30]. However, microarray technology is obsolete in comparison to state-of-the-art RNA-seq, which provides a higher dynamic range and the ability to detect novel transcripts [27, 5]. Thus, results from our study represent an important proof-of-concept that should be validated in contemporary RNA-seq cohorts, in particular datasets from The Cancer Genome Atlas (TCGA) and the TARGET initiative for pediatric leukemias [24,29]. Such validation is a necessary precondition before contemplating clinical translation.

Conclusion

In this work, we propose a dynamic integrated framework combining LightGBM and SHAP-based explainable AI for highly accurate prediction of acute leukemia subtypes. Using a standard gene expression dataset, the model demonstrated outstanding diagnostic performance (accuracy of 97.14%, AUC ROC = 0.9974), with 70 of 72 samples being correctly classified. In addition to predicting accuracy, the addition of SHAP yielded clear, clinically interpretable explanations for global and local interpretations. The authors distilled a compact genomic signature based on the framework, which is dominated by CD33 and TCF3 (well-recognized lineage markers with direct therapeutic implications), and identified nonlinear gene interactions that recaptured established regulatory networks in hematopoiesis.

Our findings illustrate that one can have biologically meaningful high-performing machine learning models. By connecting computational predictions to clinically meaningful molecular features, this method provides a template for developing other explainable AI systems in hematologic cancer. Although additional careful validation in modern RNA seq cohorts (eg, TCGA, TARGET) is needed, this work highlights the utility of explainable AI for turning genomic data into actionable diagnostic and mechanistic hypotheses, and ultimately in bolstering the ongoing paradigm shift toward precision medicine.

Limitations

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

Although the results appear to be good, there are several limitations that need to be addressed. First, the study is based on the Golub et al. (1999) microarray data set, which is a valuable reference, but the technology has been superseded by RNA-sequencing. Microarray platforms present a limited dynamic range and cannot identify novel transcripts or splicing variants [27]. Second, the sample size is small (n = 72), potentially limiting generalizability, although such concern is minimized by the use of stratified cross-validation. Third, only two acute leukemia subtypes (ALL and AML) are included in the dataset; whether the framework can be applied to rarer subtypes or for differentiating between B-ALL and T-ALL is not investigated. Fourth, while SHAP levels of interpretability are achieved, the gene interactions discovered (e.g., TCF3 and U48730_at) need further experimental studies to validate causal or functional connections. Finally, clinical translation would require prospective validation in larger, contemporary cohorts (e.g., TCGA, TARGET) to confirm robustness across varied populations and technological platforms.

References

1. Obeagu EI. Early detection and better outcomes: molecular approaches in pediatric hematologic malignancies – a review. Ann Med Surg. 2025;87(10):6564-6573. doi:10.1097/MS9.0000000000003723.
2. Singh Y, et al. Beyond post hoc explanations: a comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability. Bioengineering. 2025;12(8):879. doi:10.3390/bioengineering12080879.
3. Hoghooghi Esfahani H, Toyonaga S, Oyibo K. The application of explainable artificial intelligence in prediction, diagnosis, treatment, and management of chronic diseases: a systematic review. Digit Health. 2025;11. doi:10.1177/20552076251355669.
4. Abubakar M, et al. Spatially resolved single-cell morphometry of benign breast disease biopsy images uncovers cytomorphometric features predictive of invasive breast cancer risk. Mod Pathol. 2025;38(7):100767.
5. Genomic sequencing: techniques, advancements, and the path ahead. J Bio-X Res. 2025. doi:10.34133/jbioxresearch.0046.
6. Kumar S, Das A. Peripheral blood mononuclear cell-derived biomarker detection using explainable AI provides better diagnosis of breast cancer. Comput Biol Chem. 2023;104:107867.
7. Automated projection pursuit clustering for biological data modalities. Gigascience. 2025;14:giaf052. doi:10.1093/gigascience/giaf052.
8. Computational methods for interpretable analysis of uncertain and incomplete high-dimensional biological data. Universität Tübingen; 2025.
9. Shi Y, Huang C, Cheng H. Efficient intrusion detection system based on LightGBM algorithm. Comput Appl Softw. 2025;42(11):383-389. doi:10.3969/j.issn.1000-386x.2025.11.049.
10. Ke G, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146-3154.
11. Applications of gene pair methods in clinical research: advancing precision medicine. Mol Biomed. 2025;6:22. doi:10.1186/s43556-025-00263-w.
12. PheatPruner: interpretable feature selection for multivariate time series classification. arXiv. 2025;arXiv:2504.18329.
13. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
14. Rudin C. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-215.
15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
16. Toumaj S, Heidari A, Navimipour NJ. Leveraging explainable AI for transparent cancer detection systems. Artif Intell Med. 2025;103243.
17. Wyatt LS, et al. Explainable AI for oncological ultrasound image analysis: a systematic review. Appl Sci. 2024;14(18):8108.
18. Shin J, et al. Optimized breast cancer classification via SHAP-based feature selection. IEEE Open J Comput Soc. 2025.
19. Añez D, et al. AI pipeline for mammography-based breast cancer detection. Medicina. 2025;61(12):2237.
20. Qiu P, et al. Advancements in liquid biopsy for breast cancer. Cancer Treat Rev. 2025;139:102979.
21. Simancas-Racines D, et al. Liquid biopsy and multi-omic biomarkers in breast cancer. Biomedicines. 2025;13(12):3073.
22. Obeagu EI. Early detection and better outcomes: molecular approaches in pediatric hematologic malignancies. Ann Med Surg. 2025;87(10):6564-6573.
23. Singh Y, et al. Beyond post hoc explanations: a framework for accountable AI in medical imaging. Bioengineering. 2025;12(8):879.
24. Hoghooghi Esfahani H, et al. Explainable AI in chronic diseases: a systematic review. Digit Health. 2025;11.
25. Arber DA, et al. The 2016 revision to the WHO classification of myeloid neoplasms and acute leukemia. Blood. 2016;127(20):2391-2405.
26. Khoury JD, et al. The 5th edition of the WHO classification of haematolymphoid tumours. Leukemia. 2022;36(7):1703-1719.
27. Genomic sequencing: techniques and advancements. J Bio-X Res. 2025.
28. Kumar S, Das A. PBMC biomarker detection using XAI for breast cancer. Comput Biol Chem. 2023;104:107867.
29. Añez D, et al. AI pipeline for mammography-based cancer detection. Medicina. 2025;61(12):2237.
30. Qiu P, et al. Liquid biopsy in breast cancer: biomarkers and applications. Cancer Treat Rev. 2025;139:102979.

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*

31. Elsayed M, et al. Machine and deep learning for ALL diagnosis: a systematic review. Comput Biol Med. 2026;172:108876.
32. Shi Y, et al. Efficient intrusion detection based on LightGBM. Comput Appl Softw. 2025;42(11):383-389.
33. Ke G, et al. LightGBM: highly efficient gradient boosting decision tree. NIPS. 2017;30:3146-3154.
34. LightGBM documentation. Microsoft Corporation; 2026.
35. Applications of gene pair methods in clinical research. Mol Biomed. 2025;6:22.
36. Yelure B, et al. Predictive modelling for genetic data significance. Premier Science. 2026;PJS-25-1217.
37. Topol EJ. High-performance medicine: AI in healthcare. Nat Med. 2019;25(1):44-56.
38. Rudin C. Stop explaining black box models for high-stakes decisions. Nat Mach Intell. 2019;1:206-215.
39. Wyatt LS, et al. XAI for oncological ultrasound image analysis. Appl Sci. 2024;14(18):8108.
40. Lundberg SM, Lee SI. A unified approach to interpreting predictions. NIPS. 2017;30.
41. Parwez K, et al. Explainable federated transformer for leukemia classification. Sci Rep. 2026;16:1234.
42. Shin J, et al. SHAP-based feature selection for breast cancer classification. IEEE Open J Comput Soc. 2025.
43. Toumaj S, et al. Leveraging XAI for trustworthy cancer detection. Artif Intell Med. 2025;103243.
44. Simancas-Racines D, et al. Liquid biopsy and multi-omic biomarkers. Biomedicines. 2025;13(12):3073.
45. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer. Science. 1999;286(5439):531-537.
46. Clarke R, Ressom HW, Wang A, et al. High-dimensional data spaces. Nat Rev Cancer. 2008;8(1):37-49.
47. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. Springer; 2009.
48. Huber W, Carey VJ, Gentleman R, et al. Genomic analysis with Bioconductor. Nat Methods. 2015;12(2):115-121.
49. Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013.
50. Ioffe S, Szegedy C. Batch normalization. Proc ICML. 2015;37:448-456.
51. Kaufman S, Rosset S, Perlich C. Leakage in data mining. ACM Trans Knowl Discov Data. 2012;6(4):1-21.
52. Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.
53. Dudoit S, et al. Classification of tumors using gene expression data. J Am Stat Assoc. 2002;97:77-87.
54. Varma S, Simon R. Bias in cross-validation. BMC Bioinformatics. 2006;7:91.
55. Kohavi R. Cross-validation and bootstrap. Proc IJCAI. 1995;14:1137-1145.
56. Diamantidis NA, et al. Unsupervised stratification. Artif Intell. 2000;116:1-16.
57. Bengio Y, Grandvalet Y. No unbiased estimator for k-fold variance. J Mach Learn Res. 2004;5:1089-1105.
58. Dobbin KK, Simon RM. Splitting cases for classifiers. BMC Med Genomics. 2011;4:31.
59. Ke G, et al. LightGBM. Adv Neural Inf Process Syst. 2017;30:3146-3154.
60. Shi Y, et al. Intrusion detection using LightGBM. Comput Appl Softw. 2025;42:383-389.
61. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13:281-305.
62. Lundberg SM, Erion G, Chen H, et al. Explainable AI for trees. Nat Mach Intell. 2020;2:56-67.
63. Toumaj S, et al. XAI for cancer detection. Artif Intell Med. 2025;103243.
64. Shin J, et al. SHAP-based classification. IEEE Open J Comput Soc. 2025.
65. Wyatt LS, et al. XAI in ultrasound. Appl Sci. 2024;14:8108.
66. Saito T, Rehmsmeier M. Precision-recall vs ROC. PLoS One. 2015;10:e0118432.
67. Powers DMW. Evaluation metrics. J Mach Learn Technol. 2011;2:37-63.
68. Fawcett T. ROC analysis. Pattern Recognit Lett. 2006;27:861-874.
69. Matthews BW. Secondary structure prediction. Biochim Biophys Acta. 1975;405:442-451.
70. McKinney W. Data structures for Python. Proc Python Sci Conf. 2010;56-61.
71. Harris CR, et al. NumPy. Nature. 2020;585:357-362.
72. Hunter JD. Matplotlib. Comput Sci Eng. 2007;9:90-95.
73. Peng RD. Reproducible research. Science. 2011;334:1226-1227.
74. Lundberg SM, Erion GG, Lee SI. Consistent individualized feature attribution for tree ensembles. arXiv. 2018;arXiv:1802.03888.

Received*: 01-02-2026 -* ***Accepted****: 29-03-2026 -* ***Published****: 06-04-2026*