International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I6P120

Volume 74 Issue 6, 269-293, June 2026

Original Article

FastPedia-ML: An Interpretable Machine-Learning Framework for Pediatric Leukemia Subtype Classification using Gene-Expression Data

SAMUNDI R[1], VIJAYARANI J[2]

[1,2]*Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Padur, Tamil Nadu, India.*

[1]*Corresponding Author : rp.24cr9020002@student.hindustanuniv.ac.in*

Received: 28 December 2025 Revised: 24 March 2026 Accepted: 28 March 2026 Published: 27 June 2026

Abstract *- Pediatric Acute Myeloid Leukemia (pAML) is a heterogeneous disease with complicated genomic variants that make it difficult to subclassify the disease properly. The paper suggests a strong and explainable machine learning model applied to the classification of pediatric leukemia subtypes based on high-dimensional data in microarray gene expression. The framework combines ANOVA-based feature selection and variance-based filtering to minimize dimensionality, as well as adaptive SMOTE in order to deal with the imbalance of classes. The strategy of cross-validation is used in a nested way to guarantee the unbiased model evaluation and hyperparameters optimum. Three classifiers, which are Random Forest (RF), Support Vector Machine (SVM), and XGBoost are compared in the terms of weighted F1-score, MCC, and ROC-AUC. The experimental results on the GSE9476 dataset indicate that RF and SVM can be used to obtain perfect classification performance (F1-score = 1.000), whereas XGBoost can be used to obtain competitive results (F1 = 0.919). Statistical significance (p = 0.001) is proven by permutation testing. SHAP-based analysis also determines the biologically significant genes that correlate with the development of leukemia. The suggested framework has a high predictive power, robustness, and interpretability, which shows the possibility of using it in the context of precision medicine to diagnose pediatric leukemia.*

Keywords *- Pediatric Acute Myeloid Leukemia (Paml), Gene Expression Analysis, Microarray Data, Machine Learning, Random Forest, Support Vector Machine, Xgboost, Feature Selection, Anova, Smote, Nested Cross-Validation, High-Dimensional Data, Shap, Explainable Artificial Intelligence, Biomarker Identification, Precision Medicine*.

1. Introduction

Pediatric Acute Myeloid Leukemia (pAML) is a very non-homogenous hematological malignancy that occurs as a result of abnormal growth of immature hematopoeietic cells resulting in the disturbance of normal blood formation. Although the treatment strategies and supportive care have made significant progress, leukemia is still a significant cause of the deaths of children all over the world. Despite the positive results in overcoming the challenges of survival rate and risk stratification as well as intensive therapies, the obstacles to clinical outcome remain the recurrence, heterogeneity of the disease and difficulties in diagnosing it [2].

The last evolution of genomic technologies with fast throughput has facilitated the research of leukemia by providing the opportunity to study the molecular signatures of the disease subtypes on a large-scale scale with microarray gene expression profiling and RNA sequencing. The technologies have enabled the process of identification of unique genomic changes and patterns of expression, which has enhanced the knowledge of leukemia biology and has provided a more accurate means of classification [3]. Specifically, recent research studies have shown that pediatric AML has specific genomic features in relation to adult AML, with different molecular groups and driver mutations impacting the development of the disease and its prognosis [4].

As more and more high-dimensional genomic data has become available, Machine Learning (ML) has become a potent method of analysing complex biomedical data. The nonlinear relationships between the features of the gene expression and disease phenotypes can be captured with the help of the ML algorithms, therefore, allowing the usage of this tool to perform accurate classification, prediction of the prognosis, and discovering the biomarkers. A number of studies have been effective in using ML techniques to classify and predict leukemia with better diagnostic accuracy and predictive power [5, 6]. Moreover, ML-based systems have

demonstrated possible potential in recognizing clinically significant characteristics and assisting in treatment approaches tailored to the patient in leukemia treatment [7].

Nevertheless, in spite of such developments there are a number of limitations that exist in the current research. Numerous methods are based on standard validation methods that can bring about optimistic bias, especially were used on small and imbalanced genomic data. Moreover, the combination of feature selection, class imbalance management, and the strong validation mechanisms into one system is commonly unavailable. The second important drawback is that many ML models are not interpretable, which limits their usage in clinical practice where it is necessary to understand the role played by a single gene in the model.

An example is a linear programming-based method that was suggested in the study by Ilyas et al. [1] to classify leukemia using gene expression data which gave high accuracy with a smaller set of features. Although the study showed that the feature selection would help to enhance the classification performance, it was based on few features and applied no complex validation methods or interpretability procedures. This brings out the necessity of stronger frameworks that are able to process high-dimensional data and give credible and interpretable forecasts.

To overcome these problems, the proposed study will adopt a statistically sound and detailed machine learning model to classify the subtypes of leukemia in children based on genome-wide gene expression. The suggested method combines use of variance-based filtering, ANOVA-based feature selection, adaptive class balancing through SMOTE and nested cross-validation to be able to provide unbiased performance analysis.

Several machine learning models, such as the Random Forests, the Support Vector Machines, and XGBoost are used to represent the various learning patterns and enhance the strength of classification. Moreover, SHapley Additive exPlanations (SHAP) are also included to improve the interpretability of the models and to outline biologically meaningful gene markers that are related to leukemia subtypes.

The GSE9476 dataset is used to confirm the effectiveness of the proposed framework, high predictive performance, stability, and statistical significance. The findings confirm the hypothesis that the combination of feature selection, adaptive balancing, and nested validation denotes a substantial enhancement in model generalization and robustness in the case of high-dimensional genomic data.

The rest of the paper is structured in the following way. Section 2 gives an elaborate literature review of available methods of classifying leukemia. The third section (3) presents the proposed approach, which includes preprocessing, feature selection and model development. Section 4 talks about the performance analysis and the results of the experiment. Lastly, the study is concluded in Section 5, which also describes the future research directions.

2. Literature Review

Recent advances in Acute Myeloid Leukemia (AML) have aimed at incorporating the elements of molecular biology, genomic profiling, and machine learning in enhancing the understanding and prognosis of the disease. AML is a very heterogeneous disease with various genetic mutations and molecular abnormalities, which determine the result of treatment and survival of patients [8, 9]. Molecular research has made important developments in discovering biomarkers like CDK6 and ENO1, which are important in the progression of leukemia, cell-cycle and chemoresistance [10, 11].

Simultaneously, the technologies of machine learning have been utilized to study sophisticated genomic and clinical data more often. Such methods have demonstrated usefulness in forecasting outcomes of survival, determining prognostic variables, and encouraging precision medicine [12, 13]. Also, systems powered by artificial intelligence have been used in clinical diagnostics, including the detection of minimal residual disease, which is much more accurate and efficient than usual approaches [14]. Altogether, combining computational models with biological data contributed greatly to the study of leukemia.

More concentrated analysis of literature indicates that the current research is divided into molecular research, machine learning-based predictive models, and integrative research. Genetic mutations and biomarkers in the classification and prognosis of AML are also a focus of molecular studies, where the genes NPM1 and CDK6, among others, play critical roles in the pathophysiology of AML [9, 10]. In the same breath, metabolic controllers such as ENO1 have been found to play a pivotal role in the activity of leukemia stem cells and resistance to treatment [11].

The main approaches that are based on machine learning are aimed at predictive modelling, which is implemented with different algorithms like Random Forest, Support Vector Machine, and Gradient Boosting. These are models that have been shown to be more accurate in prediction of survival and selection of features especially in the high-dimensional datasets [12, 13]. As well, AI-based models have been successfully applied in clinical settings to identify residual disease, indicating that they can also be applied in healthcare contexts [14].

Moreover, it has been suggested to use integrative approaches using multi-omics data and machine learning to

promote better prediction and personalized treatment approaches [15, 16]. Nevertheless, most of the current research has certain drawbacks, including high dimensionality, small sample sizes, weak validation and insufficient interpretability, suggesting the necessity of more comprehensive and robust frameworks.

Table 1. Summary of related works in Pediatric leukemia classification and prognosis

S.No	Title of the Paper	Methodology Used	Inference
1	Machine learning approaches reveal methylation signatures associated with pediatric AML recurrence [17]	Boruta, LASSO, LightGBM, MCFS with Incremental Feature Selection and Random Forest classification	Identified key methylation biomarkers strongly associated with AML recurrence, improving prognostic modelling
2	Acute myeloid leukemia risk stratification using transcriptomic machine learning models [18]	RNA-seq data with k-mer based feature extraction and machine learning classification models	Achieved high accuracy (>90%) in risk prediction and identified age-related transcriptomic differences for AML prognosis
3	Exploring pattern of relapse in pediatric leukemia using machine learning methods [19]	Random Forest with clinical variables and interpretable ML (LIME)	Demonstrated effective prediction of post-transplant relapse and highlighted important clinical risk factors
4	Predicting delayed methotrexate elimination in pediatric ALL using ML [20]	LASSO feature selection with XGBoost, SMOTE, and multi-center clinical dataset	Developed a high-performance predictive model (AUROC ~0.89) for early risk detection of drug toxicity
5	Linear programming-based leukemia classification using gene expression data [1]	Feature selection with Linear Programming model on microarray dataset (GSE9476)	Achieved high classification accuracy (98.44%) for leukemia subtype prediction using reduced feature set
7	Integrating transcriptomic profiling and ML for infant AML prognosis [21]	RNA expression-based model with ML and internal/external validation	Proposed IPS score model improving risk stratification and clinical decision-making in infant AML
8	A six-gene prognostic signature for AML using machine learning [22]	Transcriptomic analysis with L0-regularized AUC optimization and risk scoring model	Identified 6-gene signature for survival prediction, improving risk stratification over traditional methods
9	Machine learning-based methylation signature analysis for AML [17]	Feature selection with Boruta, LASSO, LightGBM, MCFS and classification models	Demonstrated that methylation biomarkers significantly contribute to AML progression and recurrence prediction

The assessed articles indicate that machine learning methods have contributed immensely to the diagnosis, prognosis, and prediction of survival in leukemia as they have effectively analysed high-dimensional clinical and genomic data.

Some of the approaches have been successfully used to perform tasks like subtype classification, prediction of relapse, survival analysis, and biomarker identification [17] including Support Vector Machine, gradient boosting, and deep learning models. Also, feature selection methods include LASSO, Boruta and transcriptomic profiling and have proven useful in finding biologically significant genes and enhancing model accuracy.

Even in the face of these developments, there are still a number of serious limitations. First, numerous works are centered around clinical data, or genomic data, without an integrated framework, which effectively incorporates features selection, class imbalance management, and strong validation. Second, microarray data is associated with high-dimensionality and small sample sizes, which tend to overfit and cause unstable model performance. Third, some of the works utilize traditional validation methods, and this can be biased and compromise generalizability. Moreover, there has been a little focus on explainable artificial intelligence that constrained the biological capability to interpret predictive models.

The possible solution to these problems is the proposed research of a powerful and reproducible machine learning structure that combines variance-based filtering, ANOVA-based feature selection, adaptive SMOTE to balance the classes, and nested cross-validation as the unbiased evaluation of the performance. Besides, SHAP-based interpretability can be used to infer biologically significant aspects of genomes that close the divide between computational prediction and clinical utility.

Therefore, our methodology is set to address the weaknesses of the current methods and produce a better model by enhancing stability, generalization, and interpretability, leading to a more dependable and clinically relevant solution

to the problem of pediatric leukemia subtypes classification.

3. Research Methodology

3.1. Overview of the Proposed Framework

The present research paper hypothesizes a hypothesis whereby a statistically significant and interpretable machine learning model is reproducible and applied to the classification of subtypes of pediatric leukemia with the use of genome-wide microarray gene expression data. The datasets of Bioleukemia Pediatric leukemia have a systematic high dimensionality, small sizes, and imbalance of classes across biological subtypes. The traits usually cause overfitting and unpredictable predictive models when the traditional validation processes are employed. Thus, the suggested framework will help mitigate these issues by incorporating the dimensionality reduction, adaptive class balancing, nested cross-validation, and explainable artificial intelligence methods into a single learning pipeline.

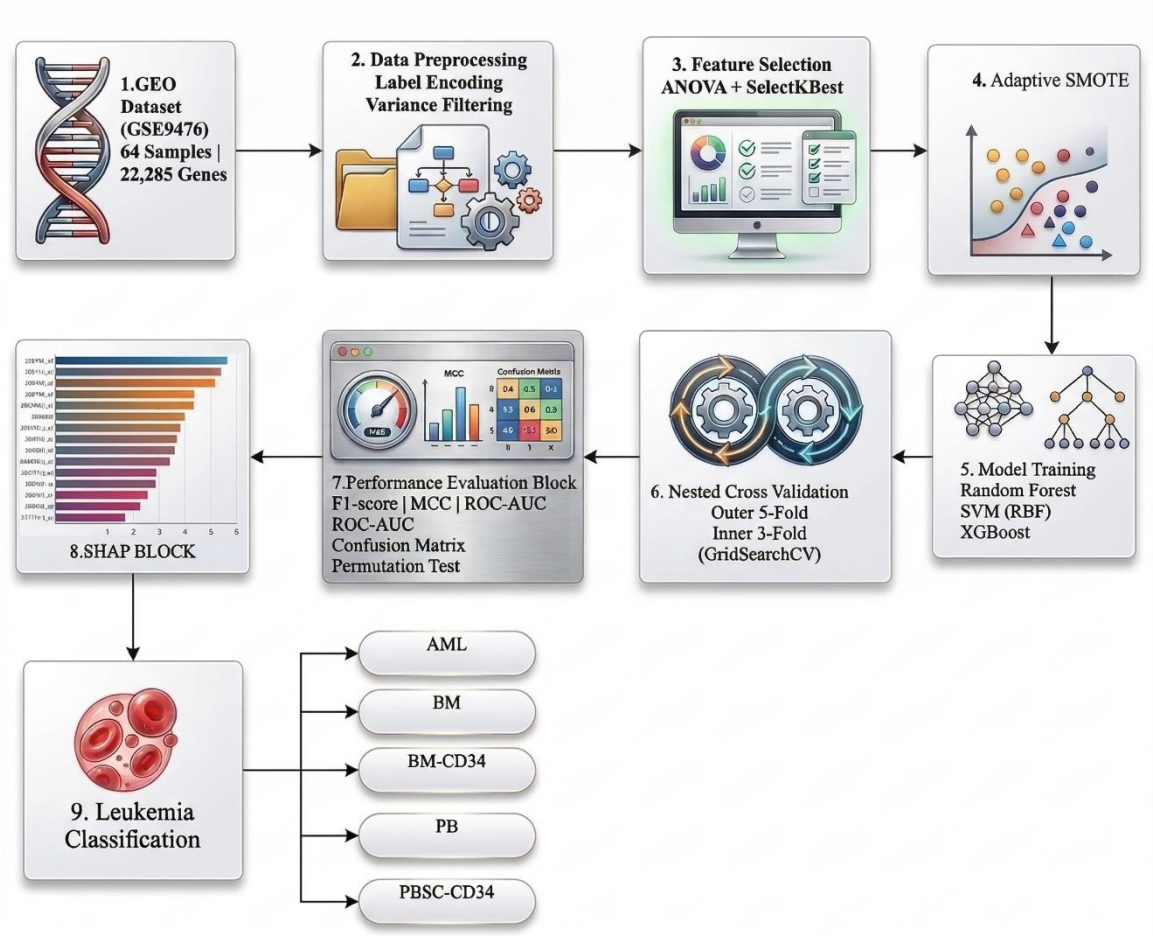

Fig. 1 Architecture of the proposed FastPedia-ML pipeline for leukemia subtype classification

The proposed pipeline will follow various sequential steps: acquiring dataset, preprocessing, filtering based on variances, statistical features selection, adaptive rebalancing of the class, training model, validating the model, and biologically interpreting the model. All stages are applied in one machine learning pipeline to avoid data leakage and to have an objective evaluation of the models. The process starts by first obtaining the microarray gene-expression data in the Gene Expression Omnibus repository, and preprocessing tasks like encoding of labels and filtering of variances are performed to eliminate redundant or non-informative gene probes. The elimination of genes whose variance is almost zero can help eliminate noise and enhance the computational speed in the context of high-dimensional genomic data. Informative genes are then selected after filtering with a feature selection method based on Analysis Of Variance (ANOVA) with the SelectKBest algorithm. The method assesses the statistical relationship between gene expression measures and the labels of the leukemia subtypes, which enables the presentation of gene expressions that have a pronounced discriminative ability across the classes.

The selected genes are a hyperparameter to optimize the dimensionality of the feature space during model training to ensure that its size is optimal. This strategy will make sure that the feature subset that results maintains the most biologically

informative signals and minimizes the chances of overfitting. As the dataset is not equally distributed among leukemia subtypes, the problem of class imbalance is resolved with a Syntactic Minorities Oversampling Technique (SMOTE) in an adaptive manner. SMOTE uses interpolation between the neighbouring samples in the feature space to create synthetic samples of minority classes. To ensure stability in small datasets, the number of nearest neighbours to be used in the process of oversampling is dynamically set to the smallest class size within each training fold. Notably, the oversampling is only used on the training folds of the cross-validation exercise and hence the test data does not leak information.

After feature selection and class balancing, three supervised machine learning algorithms are trained and tested, namely, the Random Forest (RF), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). Such algorithms are complementary learning paradigms such as ensemble bagging, kernel-based classification and gradient boosting respectively. The framework intends to select the best classifier that will be the most robust in predicting pediatric leukemia subtypes by comparing several model families.

The framework uses nested cross-validation to achieve accurate performance estimates and to prevent the existence of optimistic bias. Outer cross-validation loop is applied to the evaluation of performance unbiasedly and the inner loop is applied to the optimization of hyperparameters with the help of the GridSearchCV.

The two-level validation plan is used to make sure that hyperparameter optimization is done solely on training data, which avoids information leakage and gives a more realistic prediction of the ability of the model to generalize.

Several independent metrics are used to measure model performance, which are weighted F1-score, Matthews Correlation Coefficient (MCC), and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). The weighted F1-score of the outer cross-validation folds is assumed to be the most important measure of performance since it balances the precision and the recall of all the leukemia types. Alongside, there is also permutation testing to determine the statistical significance of model performance and also to ensure that the observed classification accuracy is not realized by mere chance.

Lastly, SHapley Additive exPlanations (SHAP) is used to analyze trained models in order to enhance interpretability and connect computational predictions to biological understanding. SHAP values predict the amount that each gene contributes to the model predictions and allow the identification of biologically meaningful genomic markers with leukemia subtypes. The explainable artificial intelligence allows the framework to achieve high predictive accuracy in addition to providing meaningful insight into the molecular mechanisms of pediatric leukemia.

The general scheme of the proposed machine learning structure can be depicted in Figure 1, which is a summary of the entire workflow of data preprocessing to model interpretation.

Algorithm 1: FastPedia-ML Pipeline for Pediatric Leukemia Subtype Classification
Input: Gene expression dataset (GSE9476)
Output: Predicted leukemia subtype + important genes

Step 1: Load gene expression dataset from GEO repository.
Step 2: Convert leukemia subtype labels into numeric values using label encoding.
Step 3: Remove genes with zero or very low variance (non-informative features).
Step 4: Apply ANOVA-based feature selection (SelectKBest) to choose top important genes (e.g., 200 features).
Step 5: Handle class imbalance using adaptive SMOTE on training data only.
Step 6: Train machine learning classifiers: Random Forest (RF), Support Vector Machine (SVM), and XGBoost.
Step 7: Perform nested cross-validation:

- Inner loop → hyperparameter tuning
- Outer loop → model evaluation

Step 8: Evaluate performance using: Accuracy, Precision, Recall, F1-score, MCC, ROC-AUC)
Step 9: Perform permutation testing to verify statistical significance.
Step 10: Apply SHAP analysis to identify important genes contributing to classification.

Output: Best classifier predicts leukemia subtype and identifies key biomarker genes.

3.2. Dataset Description

The GSE9476 pediatric leukemia microarray dataset published in the Gene Expression Omnibus (GEO) in the public repository at the National Center of Biotechnology Information (NCBI) [23] was used as experimental data in the course of this study. The data format comprises genome-wide gene expression patterns of pediatric leukemia samples and associated hematopoietic cell populations, and thus it is appropriate to study and determine the relationship between molecular signatures and leukemia subtype differentiation. The dataset is composed of 64 biological samples, and it has about 22,285 gene probes, which are an expression measure derived by microarray technology. The probes are designed to correspond to the gene transcripts with the levels of expression indicating the level of the biological activity of the respective gene in the sample under analysis. The samples are divided into five biologically different groups of various hematopoietic states and subtypes of leukemia:

- Acute Myeloid Leukemia (AML)
- Bone Marrow (BM)
- Bone Marrow CD34 (BM-CD34)
- Peripheral Blood (PB)
- Peripheral Blood Stem Cell CD34 (PBSC-CD34).

These classes are malignant and normal populations of hematopoietic cells that make it possible to create a classification model that is capable of differentiating not only between the samples of leukemia and normal cellular populations but also distinguishing between biological subtypes related to each other.

There is a moderate level of imbalance in the distribution of the samples among the five classes. Specifically, the number of samples available in the AML subtype is the greatest, whereas some populations of CD34-enriched have fewer observations. This imbalance of classes may create biasness in the predictive models unless it is well tackled in the learning process. Label encoding Before model training, the labels of the categorical subtypes were transformed to numeric values so that they could be compatible with machine learning algorithms. Once the encoding is finished, the dataset may be expressed as a high dimensional feature matrix.

$$X \in \mathbb{R}^{n \times d} \quad (1)$$

and the number of samples is represented by n=64 and number of features of gene expression is represented by d=22,285. The target vector y has the encoded subtype labels of each sample.

One of the most significant challenges related to this dataset is that it has a very large ratio of features to samples in which the number of genes is significantly larger than the number of observations. This high dimensionality may cause various computational and statistical problems, such as the curse of dimensionality, higher variability in model estimation, and higher chance of overfitting. Moreover, a high number of microarray probes can be of little or weak variation or association with leukemia subtypes, which will introduce noise into the learning process.

To overcome these problems, the planned framework will use the variance-based filtering, statistical feature selection, and nested cross-validation to eliminate dimensions without any loss of information about biologically relevant factors. The following steps will mean that the resulting predictive models will be robust and generalizable even when the sample size is small as will be the case with pediatric leukemia genomic data.

3.3. Problem Formulation

The main goal of the study is to create a machine learning model that will be able to predict the pediatric leukemia subtypes with a high degree of accuracy with the help of the high-dimensional gene expression data. Resting on a collection of microarray samples, one can define the problem as a multi-class supervised classification problem where each of the samples is modeled by a collection of gene-expression values and has a biological subtype label attached. Officially, the mapping task can be defined as a mapping function.

$$f: X \rightarrow Y \quad (2)$$

Where X represents the gene expression feature matrix and Y denotes the set of leukemia subtype labels. Each sample $x_i \in \mathbb{R}^d$ corresponds to a gene expression vector containing measurements for thousands of genes, where d denotes the number of gene probes. In this study, $d = 22{,}285$, representing the total number of microarray probes in the dataset. The target label y which is assigned to each sample is one of a finite set of biologically related leukemia categories:

$$y \in \{AML, BM, BM\_CD34, PB, PBSC\_CD34\} \quad (3)$$

Where AML denotes Acute Myeloid Leukemia, BM represents Bone Marrow cells, BM-CD34 corresponds to CD34-enriched bone marrow stem cells, PB indicates Peripheral Blood samples, and PBSC-CD34 refers to Peripheral Blood Stem Cell CD34 populations. The goal of the predictive model is to learn a decision function $f(x)$ that maps each gene expression profile to its correct subtype label. Although, there are a number of major concerns with predictive modeling of genomic data.

3.3.1. High Dimensionality

The size of the gene features (d=22,285) is extremely huge in comparison with the number of available samples (n=64). This feature to sample ratio is the source of the so-called curse of dimensionality where the feature space is sparse and learning algorithms based on distance cannot easily generalize. In the absence of dimensionality reduction, it is possible that models will be picking up noise instead of biological patterns of significance.

3.3.2. Small Sample Size

Biomedical data may have small samples that are labeled, as it is challenging to gather clinical data. The danger of overfitting is high since there are only sixty-four available samples. Unless sound validation exercises are adopted, a model can be effective on training data, but it may fail when applied on unknown samples.

3.3.3. Class Imbalance

The sampling of the leukemia subtypes is not equalized. The sample of the AML class is the highest, and certain classes associated with stem cells have fewer samples. This imbalance may cause classifiers to be biased with majority classes; minor subtypes have less predictive power.

3.3.4. Biological Noise and Redundancy

Microarray data usually include thousands of gene probes which can be weakly associated or not related to the disease phenotype. Such redundant or noisy factors may hide useful patterns and have adverse impacts on model stability and interpretability.

In order to solve these problems, several methodological parts are included in the offered framework. Filtering is done to eliminate non-informative genes using variance-based filtering and ANOVA-based feature selection is then done to identify the discriminative genomic features. The imbalance between classes is reduced by applying an adaptive SMOTE technique that creates synthetic samples of minority groups within the training folds. Lastly, nested cross-validation is used to guarantee an objective assessment of the models and avoid information leakage on hyperparameter optimization.

The proposed framework will utilize dimensionality reduction, adaptive class balancing, and powerful validation methods to learn a sound mapping function that can provide high accuracy in classifying these subtypes of pediatric leukemia in terms of high-dimensional gene expression data.

3.4. Data Preprocessing

Preprocessing was done to guarantee the numerical consistency of the data, eliminate non-informative gene probes, and to prepare the data to undergo machine learning analysis in a reliable way. Taking into account such a high dimensionality of microarray gene expression data, preprocessing is of paramount importance in enhancing model stability and decreasing the role of noisy or redundant features. All preprocessing was done as a part of a machine learning pipeline so that transformations were only done on training data when cross-validation was done, and thus information leakage was avoided.

The initial preprocessing involved the label encoding of categorical labels of leukemia subtypes and converting them into numbers. This transformation is a set of biological classes receiving a unique integer value, which the machine learning algorithms are able to process the target variable. The coded labels are the five leukemia related categories: Acute Myeloid Leukemia (AML), Bone Marrow (BM), Bone Marrow CD34 (BM-CD34), Peripheral Blood (PB), and Peripheral Blood Stem Cell CD34 (PBSC-CD34).

In order to eliminate redundant features and noise, a variance-based filtering method was used. A gene probe that has a zero or near zero variance between samples does not provide much or no information to use in the classification task. These characteristics add little to the learning experience and can raise the computational barrier and not predictive performance. Thus, the elements that had a zero variance were filtered out with a variance threshold filter.

The variance of each gene probe was calculated as

$$Var(X_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \quad (4)$$

Where x_{ij} represents the expression value of gene j in sample i, n denotes the number of samples, and $\bar{x}_j$ represents the mean expression value of gene j across all samples. Attributes whose variance was equal to zero were then eliminated, thus cutting down on the dimensions and enhancing the computing power but not eliminating informative genes.

Besides the feature filtering, feature scaling was also used in those algorithms that were sensitive to the variability in the magnitude of the features. Specifically, Support Vector Machines are based on distance-based kernel functions, and they are sensitive to the extent of input features. In order to have the same contribution of all the genes in the distance computation, we performed feature normalization through standardization.

Each feature was transformed according to

$$z = \frac{x-\mu}{\sigma} \quad (5)$$

Where x represents the original gene expression value, μ is the mean of the feature calculated from the training data, and σ denotes the corresponding standard deviation. This transformation rescales features to have zero mean and unit variance, thereby preventing features with large numerical ranges from dominating the learning process.

It should be mentioned that normalization was used on the Support Vector Machine pipeline only, as tree-based models (Random Forest and XGBoost) are automatically resistant to feature scaling because of their decision-tree splitting algorithms. As a result, these models did not include scaling so that their computational efficiency would not compromise model performance.

It was performed within the training folds of the nested cross-validation process to execute all of the preprocessing operations such as variance filtering, feature scaling. This method guarantees that preprocessing parameters are only learned based on training samples and applied to the respective test folds hence the integrity of the evaluation process is upheld and no unwanted leakage of information in the training stage to the test folds.

The proposed framework lowers the number of redundant features, ensures numerical stability, and prepares the data with an efficient feature selection and model training in the next phases of the pipeline through this form of thoughtful preprocessing strategy.

3.5. Feature Selection

The microarray data utilized in this research has over twenty thousand probes of genes whereas the sample size of the samples is insufficient. This level of feature to sample ratio can severely risk overfitting the models as well as worsen the generalization capacity of machine learning models. Consequently, it is necessary to identify a number of informative genes that would ensure a higher level of predictability and increase the interpretability of the model.

In order to solve this problem, a statistical feature selection model, which is an Analysis Of Variance (ANOVA), was used. Particularly, SelectKBest algorithm was employed to rank the gene features based on their discriminatory ability in the various subtypes of leukemia. ANOVA F-statistic is the ratio between between-class to within-class variance of each of the gene features. Features that have high F-statistic value suggest that there is stronger difference in expression levels among classes and thus they have more information to use in classification.

The F-statistic used for feature ranking is defined as,

$$F = \frac{\text{variance between classes}}{\text{variance within classes}} \tag{6}$$

Where T (i, j) is the difference between the means of the features of the various classes and V (i, j) is the difference between the value of the features of the various classes. The genes that have a higher F-score have better capabilities of class separation and are thus given precedence during the feature selection process.

Following the calculation of F-statistics of all gene probes, SelectKBest algorithm keeps the best kfeatures with the highest score. To avoid picking a set number of genes, the feature retention number was a hyperparameter to be optimized in model training. The parameter kwere had the following candidate values that were defined as:

$$k \in \{200,500,1000\} \tag{7}$$

The selection of these values was aimed at compromising dimensionality reduction to the maintenance of the biologically significant signal of gene expression.

In order to avoid information leakage, as well as to ensure model evaluation integrity, the feature selection was conducted in the training folds of the nested cross-validation framework. The ANOVA F-statistics were calculated on training data only in each training fold and the selected features were used to transform the training and validation subsets of the fold. This process gives an assurance that the test data will not be visible at all when the process of selecting features takes place.

The proposed feature selection strategy is effective in that it can reduce the dimensions of the gene expression dataset by using statistical ranking and optimization using cross-validation to retain the most discriminative genomic features. Not only does this boost the computational efficiency, but also increases both the stability and readability of the machine learning models that are used to classify the subtypes of pediatric leukemia.

3.6. Class Imbalance Handling

The pediatric leukemia dataset utilized in this paper has disproportionate sample distribution between the biological subtypes, with Acute Myeloid Leukemia (AML) class having a higher number of samples than a number of CD34-enriched hematopoietic populations. Imbalance of classes may have an adverse effect on learning process, whereby a biasness in classifier to majority classes is experienced, thus lowering the predictive ability to minority classes. The problem has been especially severe in biomedical classification problems in which it is necessary to obtain precise identification of the infrequent subtypes to obtain any credible disease characterization.

Synthetic Minority Oversampling Technique (SMOTE) was used in order to overcome this difficulty. SMOTE: This is an oversampling technique which is popular and is created to generate artificial samples of minority classes, based on interpolating between existing samples in the feature space. In contrast to mere imitation of the minor instances, SMOTE finds new samples on the line between a sample and its closest neighbours, which results in greater diversity of classes and better generalization of the classifier. The generation of a synthetic sample is defined as

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \tag{8}$$

Where x_i represents a minority class sample, x_{nn} denotes one of its nearest neighbors belonging to the same class, and $\lambda \in (0,1)$ is a randomly generated interpolation coefficient. This procedure is practical in generating a fresh sample between the already existing samples in the feature space, which broadens the minority class distribution.

Given that the dataset has a small sample size, using more traditional SMOTE and specifying the number of nearest neighbours can be unstable in case the minority group only has a small number of observations. In order to overcome this drawback, an adaptive SMOTE approach was applied. Under this method, the size of the nearest neighbours employed in the synthetic sample generation is adaptively determined with respect to the smallest of all the classes in each training fold. In particular, the value of k used by SMOTE is determined by:

$$k = \min(k_{max}, n_{minority} - 1) \tag{9}$$

Where $n_{minority}$ represents the number of samples in the smallest class within the current training fold and k_{max} denotes the maximum allowed number of neighbors. This adaptive mechanism ensures that SMOTE never attempts to use more neighbors than available samples, thereby preventing runtime errors and maintaining stable oversampling behavior for small biomedical datasets.

The other important issue of the proposed framework is the incorporation of SMOTE into machine learning pipeline. The training data of each cross-validation fold were only oversampled and thus only synthetic samples were created by using the training observations. The validation or test folds were not in any way manipulated in the oversampling process. This design averts information leaking between training and testing phases which is a prevalent causative factor of artificially enhanced performance in genomic classification research.

The proposed solution ensures the balance between the number of classes, using adaptive oversampling, and data isolation, using strict folds, with maintaining the integrity of the evaluation process. Such an approach allows the machine learning models to acquire more representative decision boundaries and enhances their capability to identify the minority leukemia subtypes accurately.

3.7. Model Development

Three machine learning algorithms Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost)) were trained on three datasets to assess the ability of the proposed framework to identify the most effective in the classification of pediatric leukemia subtypes. These algorithms are complementary learning programs, such as ensemble bagging, kernel based classification, and gradient boosting. Working with several model families enables the overall comparison of the classification performance in the situation of high-dimensional genomic conditions.

3.7.1. Random Forest Classifier

Random Forest is an ensemble learning algorithm that relies on the bagging principle according to which several decision trees are built on bootstrapped subsets of the training data. The trees are trained on a case-by-case basis and the resulting decision is reached after majority voting by all the trees in the ensemble.

The prediction function of the Random Forest classifier can be expressed as

$$f(x) = \text{majority}(T_1(x), T_2(x), \ldots, T_m(x)) \quad (10)$$

Where $T_i(x)$ denotes the prediction of the i^{th} decision tree and m represents the total number of trees in the ensemble. During tree construction, node splitting is guided by the Gini impurity criterion, which measures the degree of class heterogeneity within a node. The Gini impurity is defined as

$$G = 1 - \sum_{i=1}^{C} p_i^2 \quad (11)$$

Random Forest is one algorithm that is especially appropriate in high-dimensional data since it is capable of dealing with a significant number of features and reduces overfitting due to ensemble averaging.

3.7.2. Support Vector Machine

The Support Vector Machines are effective supervised learning algorithms, which are aimed at discovering the optimal separating hyperplane that maximizes the margin between classes. In nonlinearly separable data, like gene expression profiles, data is mapped to higher-dimensional feature spaces by the use of kernel functions which are then used to separate the data linearly.

The Radial Basis Function (RBF) kernel was used in this study and it is defined as

$$K(x_i, x_j) = \exp(-\gamma \mid\mid x_i - x_j \mid\mid^2) \quad (12)$$

Where x_i and x_j represent feature vectors, $\mid\mid x_i - x_j \mid\mid$ denotes the Euclidean distance between samples, and γ is a kernel parameter controlling the influence of individual training samples.

The SVM optimization problem aims to minimize

$$\frac{1}{2} \mid\mid w \mid\mid^2 + C\sum \xi_i \quad (13)$$

Where w represents the weight vector defining the decision boundary, C controls the trade-off between maximizing the margin and minimizing classification error, and ξ_i denotes slack variables allowing misclassification of certain samples.

Due to its reliance on distance-based kernels, feature scaling was applied specifically to the SVM pipeline to ensure numerical stability.

3.7.3. XGBoost Classifier

XGBoost is a more advanced method of gradient boosting that uses sequential decision tree building to reduce errors in prediction of past iterations. In contrast to bagging-based methods, boosting algorithms allow to enhance the performance of the model progressively by working on the hard-to-predict samples.

The objective function optimized by XGBoost is defined as:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \Omega(f) \quad (14)$$

Where $l(y_i, \hat{y}_i)$ represents the loss function measuring prediction error, and $\Omega(f)$ denotes a regularization term used to penalize model complexity.

The regularization term is given by,

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \, || \, w \, ||^2 \quad (15)$$

Where T represents the number of leaves in the tree, w denotes leaf weights, and γ and λ are regularization parameters controlling tree complexity and preventing overfitting.

XGBoost offers several advantages for genomic classification tasks, including efficient handling of high-dimensional data, built-in regularization mechanisms, and strong predictive performance.

3.7.4. Model Complementarity

The selected classifiers provide complementary perspectives on the leukemia classification problem. Random Forest models nonlinear interactions by bagging the ensemble, SVM by modeling the complex boundaries of decisions by using a kernel transformation and XGBoost trains by using gradient boosting to iteratively trim the predictions. The proposed framework can determine the strongest model to predict the subtype of leukemia in children by comparing several algorithms on a single experimental pipeline.

3.8. Nested Cross-Validation and Hyperparameter Optimization

A nested cross-validation strategy was used to get an unbiased estimation of the model performance and to eliminate optimistic biasing in the process of hyperparameter tuning. Nested cross-validation is considered to be among the most trustworthy assessment techniques to machine learning models which are trained on small and high-dimensional data, including microarray gene expression data.

The nested cross-validation process is based on two validation loops: an outer loop is the model evaluation loop, and an inner loop is the hyperparameter optimization loop. The outer loop in this study was 5-fold stratified cross-validation and the inner loop was 3-fold stratified cross-validation.

The dataset was split in 5 equally sized folds in the outer loop, and the distribution of classes among leukemia subtypes was maintained. Each iteration was trained on four folds and the other on a test set. The process was performed five times such that each fold acted as independent test set. An average of the performance measures of these outer folds was taken to offer a sound value of model generalization performance.

An inner cross-validation loop was carried out to identify the best hyperparameters to each classifier in each of the outer training fold. The grid search algorithm (GridSearchCV) was used to perform tuning of hyperparameters in a systematic way by trying out different combinations of possible values of the hyperparameters. The model was trained and cross-validated, on each candidate configuration, on 3-fold stratified cross-validation of the training data of the outer fold. The best hyperparameter configuration, based on the weighted F1-score during the inner validation process, was chosen as the desired configuration.

The nested cross-validation architecture will comprise two levels: an outer loop and an inner loop to guarantee the solid model evaluation and the best hyperparameter adjustment. The 5-fold cross-validation strategy is used in the outer loop to split the dataset into training and testing subsets to estimate the performance of the strategy in an unbiased way. Each outer training fold has an inner loop that is done in 3-fold cross-validation to do hyperparameter optimization via GridSearchCV. This methodically tests parameter combinations and picks hyperparameters with the best performance. This model is trained on the best parameters on the training part of the outer fold and tested on the relevant outer test fold to get a stable estimate of model performance.

This two-level validation scheme is instrumental in the sense that hyperparameter optimization is conducted only on their respective training sets and the outer external test folds are never visible to the optimization process. Consequently, the reported performance measures are more representative of the true predictive ability of the models and not optimistic measures obtained through overfitting.

The preprocessing and feature selection pipeline was coupled with the nested cross-validation system, which means that the variance filtering process, the feature selection process based on ANOVA, and the adaptive SMOTE oversampling process were independently conducted on each of the training folds. This kind of definite separation of the training and testing stages prevents data leaks and ensures integrity of the experimental test.

Once the technique of using nested cross-validation and systematic hyperparameter optimization has been incorporated, the proposed framework provides a statistically valid evaluation of the performance of the models in the classification of the subtypes of pediatric leukemia using the high-dimensional gene expression data.

3.9. Evaluation Metrics

The effectiveness of the suggested machine learning models was measured with the help of several complementary measures to guarantee full estimation of the classification quality. Because biomedical datasets are usually imbalanced in classes and have complicated decision boundaries, then

using one metric (e.g. accuracy) can lead to erroneous conclusions. As such, various metrics of evaluation were used such as Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC) and Receiver Operating Characteristic Area Under the Curve (ROC-AUC).

Accuracy measures the overall proportion of correctly classified samples and is defined as,

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

Where TP represents true positives, TN denotes true negatives, FP corresponds to false positives, and FN indicates false negatives. Although accuracy provides a general estimate of classification correctness, it may not adequately reflect model performance when class distributions are imbalanced.

To capture class-specific predictive performance, precision and recall were also computed. Precision measures the proportion of predicted positive samples that are correctly classified and is defined as,

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

While recall (also referred to as sensitivity) measures the proportion of actual positive samples that are correctly identified by the model:

$$Recall = \frac{TP}{TP+FN} \quad (18)$$

To provide a balanced measure of precision and recall, the F1-score was calculated as the harmonic mean of the two metrics:

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (19)$$

The weighted F1-score has been used as the key performance measure since there are several leukemia subtypes and they have different proportions in the dataset. The weighted F1-score does not ignore the class imbalance as it uses the sample size of each category of the leukemia to weight the contribution of each category and hence give a more valid measure of the overall model performance in all the categories of leukemia.

Besides these common measures, Matthews Correlation Coefficient (MCC) was also employed as a strong measure of classification performance. The imbalanced datasets that MCC is most effective with involve that it takes into account all the factors of the confusion matrix and ensures that the assessment remains balanced even in cases where the sizes of the classes are unequal. The MCC is defined as,

$$MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (20)$$

The MCC is a value between -1 and +1 where +1 will be the result of perfect classification, 0 the result of random classification and -1 the result of complete disagreement between the prediction and the true classification.

In order to further assess model discrimination ability, Receiver Operating Characteristic Area Under the Curve (ROC-AUC) was also calculated based on a One-Vs-Rest (OVR) multi-class classification strategy. ROC-AUC evaluates how well the classifier can differentiate among classes at different levels of decision threshold and gives a threshold independent measure of model performance.

Lastly, permutation test was done to confirm that the classification performance witnessed was not due to the random chance. In this process, classification labels were randomly interchanged and repetition of the process was done several times. We did 1000 permutations with the same cross-validation setting in this study. This p-value is the likelihood that the performance could happen with random assignments of labels. The p-value of less than 0.05 represents statistically significant model performance.

The metrics of all the evaluations were calculated with the help of predictions that were made by outer folds of a nested cross-validation process and the performance estimates were based on the actual generalization performance of the model.

3.10. Model Interpretability

The interpretability of machine learning in biomedical studies is a necessary attribute in cases where predictive models are applied in the investigation of possible molecular biomarkers related to disease disease mechanisms. In genomic research, it is of importance to know which genes have the most significant impact on classification choices to be able to convert computational predictions into biologically useful information. In order to accomplish this, SHapley Additive exPlanations (SHAP) were used to explain the predictive behaviour of the trained machine learning models.

SHAP is an explainability model that adopts the cooperative game theory to estimate the importance of each feature to a model prediction. SHAP framework uses the Shapley value, which is an importance score, to assign to each feature through the calculation of the marginal contribution of the feature to the prediction in all possible feature subsets. This enables the model output to be represented as an additive contribution of each of the features.

The SHAP value for a feature i is defined as,

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S\cup\{i\})-f(S)] \quad (21)$$

Where F represents the full set of features, S denotes a subset of features excluding feature i, and $f(S)$ represents the model prediction based on the subset S. The coefficient term ensures fair attribution by considering all possible combinations of feature interactions.

SHAP analysis of final model which was trained on the whole data was done in this study after the nested cross-validation assessment. With tree-based models, including Random Forest and XGBoost, a TreeExplainer implementation of SHAP was utilized and is effective to provide exact Shapley values on ensemble tree models. In the case of kernel-based models, including Support Vector Machines, KernelExplainer method was used, which computes the Shapley values via approximation on background samples of the dataset.

SHAP framework enjoys global and local interpretability. Global interpretation reveals the genes that are repeatedly used to make classification choices throughout the dataset, and this allows key genomic markers to be found that are related to the leukemia subtypes. Local interpretation is the one that is used to explain individual predictions by quantifying the effect of individual genes on the classification of a given sample.

The techniques of visualisation that were applied to determine the most significant genes in the predilection of leukemia subtypes included SHAP summary plot and feature importance rankings. The scores of gene importance that have resulted are good indications of the molecular pathways underlying the pediatric leukemia and contribute to the biological relevance of the predictive models.

The study does not only reach precise classification of the subtypes of leukemia, but also increases interpretability of the machine learning prediction, which makes it possible to identify biologically significant genomic characteristics related to progression of pediatric leukemia by incorporating SHAP-based explainability in the offered framework.

4. Results and Analysis

In this section, the experimental analysis of the suggested machine learning model to classify genome-wide microarray gene expression of pediatric leukemia subtypes is performed. The data is analyzed in the order of the pipeline outlined in the methodology, i.e. it starts with characterizing data and reducing features, then correcting class imbalances, training a model, and validating the model with the help of the nested cross-validation. There are a number of complementary measures that are used to determine the predictive performance of the multiple classifiers such as weighted F1-score, Matthews Correlation Coefficient (MCC) and ROC-AUC. Moreover, statistical validation by means of permutation testing, interpretability analysis by SHAP is conducted to guarantee the soundness and biological significance of findings. The results show that the proposed pipeline is capable of processing high-dimensional genomic data on small samples and correctly determining discriminative features of genes related to pediatric leukemia subtypes.

4.1. Dataset Characteristics

An experimental analysis of the suggested classification framework was performed on the GSE9476 pediatric leukemia microarray dataset, which was retrieved through the Gene Expression Omnibus (GEO) repository. It is a complete set of genome-wide gene expression data or profiles obtained by microarray technology, which is widely applied in biomedical studies in the study of transcriptional patterns of leukemia.

The data is in the form of 64 samples and about 22,285 gene expression probes with a probe referencing the level of expression of a given gene or transcript in all the samples collected. These samples are part of five biologically diverse groups linked with leukemia and normal population of hematopoietic cells:

- Acute Myeloid Leukemia (AML)
- Bone Marrow (BM)
- Bone Marrow CD34 (BM-CD34)
- Peripheral Blood (PB)
- Peripheral 34 blood stem cell CD34 (PBSC-CD34)

These are the various types of hematopoietic cells and various stages of its development in the blood and bone marrow system. The fact that both leukemia samples and normal populations of hematopoietic cells are introduced makes it possible to create machine learning models that could identify the difference between malignant cells and healthy cellular conditions.

Table 1. Sample distribution of leukemia subtypes

Class	Samples
Acute Myeloid Leukemia (AML)	26
Bone Marrow (BM)	10
Bone Marrow CD34	8
Peripheral Blood (PB)	10
Peripheral Blood Stem Cell CD34	10
Total	64

The dataset does not have an extreme class imbalance as the AML subtype has the most significant share of samples and the Bone Marrow CD34 category represents the fewest group as indicated in Table 1. The issue of class imbalance is typical of biomedical data: some types of clinical samples are not easily accessible. Otherwise, such imbalance can lead to machine learning algorithms having a bias in favor of the majority classes, and also lower the predictive performance on the minority subtypes.

The other serious problem that comes with this dataset is that it has very high dimensionality. The dataset has over 22,000 features of gene expressions whereas there are only 64 samples available. This imbalance between features and observations is often known as the curse of dimensionality. In this case, machine learning models can fit the training data in a way that memorizes noise instead of learning any significant biological patterns.

Large high-dimensional genomic datasets have also a high count of redundant or weakly informative gene probes that may adversely impact the model and raise the amount of computation. Thus, it is necessary to have effective feature reduction methods that help in enhancing the generalization of a model and to identify biologically relevant gene signatures.

Nevertheless, GSE9476 dataset is a good reference on how machine learning methods can be applied in the classification of the pediatric leukemia subtype. Molecular signatures in the progression of leukemia can be identified using its genome-wide gene expression measurements and used to develop predictive models that could aid in the early diagnosis and precision medicine of pediatric oncology.

4.2. Feature Reduction Results

GSE9476 dataset has about 22,285 gene expression probes which makes machine learning models very challenging to train. In the cases where features are far more than the samples, models can be afflicted by the curse of dimensionality and become overfitted, with high variance and low predictive power on unobserved data. Consequently, the key to success in dimensionality reduction strategy is to identify meaningful genomic signals and to get rid of redundant or noisey features.

In order to solve this problem, a two-step feature reduction pipeline was introduced in the proposed structure. The former stage used Variance Threshold filtering, whereas the latter used ANOVA-based statistical feature selection on SelectKBest method.

Variance Threshold filtering was applied in the initial step to filter off gene probes that have zero or close to zero variance across the samples. These features are nearly fixed across all observations and hence cannot be of use in the task of classification. Eliminating these features assists in getting rid of the redundant probes and diminishing the dimensions of the dataset without losing any valuable patterns of gene expression.

SelectKBest feature selection through ANOVA was then used as the second step of dimensional reduction after the variance filtering. It is a statistical tool, which assesses every gene characteristic by F-statistic of ANOVA, which is a ratio between between-class and within-class variances. Attributes with larger F-scores provide stronger evidence of difference in gene expression amongst leukemia sub-types and hence would be more valuable in classification.

The parameter k, which is the number of retained genes, was a hyperparameter to be optimized in the course of the nested cross-validation to define what is the best number of selected features. Three values of candidates were tested:

$$k \in \{200,500,1000\}$$

In the inner cross-validation loop, each candidate feature subset was trained to obtain models, and the one reporting the most weighted F1-score was then chosen. This tuning step would make sure that the feature selection phase would be kept as part of the training pipeline, and information leakage of the testing data will not occur.

The experiment found that choosing about 200 features of the genes gave the best and most stable classification performance. Adding more features than this point did not result in major changes in predictive performance and sometimes also added more noise to the model.

The step of decreasing the number of more than 22,000 probes on the gene array to about 200 informative genes offered a number of significant benefits. To begin with, it dramatically decreased the computational complexity and the time to train. Second, it removed many redundant and weakly informative gene probes, and hence increased noise in the dataset. Lastly, the feature space, which was reduced, also enhanced the stability and generalization of the models and enabled the classifiers to specialize in biologically relevant patterns of gene expression related to pediatric leukemia subtypes.

Altogether, the variance-based filtering in conjunction with ANOVA-based feature selection was a powerful approach to reducing the massive-dimensional genomic data to a small and informative display of features that can be used in machine learning classification.

4.3. Class Distribution before and after SMOTE

The GSE9476 dataset has average class imbalance of leukemia subtypes which is typical feature of biomedical datasets. In skewed data sets, machine learning systems are more likely to learn majority classes and make biased predictions, as well as, become less sensitive to minority classes. This is of significant concern when it comes to medical use since proper diagnosis of the sub-type of minority diseases is of paramount importance.

Table 2 shows the initial distribution of classes of the data set prior to the use of any class balancing method. The class of Acute Myeloid Leukemia (AML) has the greatest amount of samples and the Bone Marrow CD34 (BM-CD34) group is

the smallest as indicated in Table 2. This type of imbalance can lead to classification models paying more emphasis to the majority AML class, but it performs poorly in recognition of minority hematopoietic subtypes.

Table 2. Class distribution before SMOTE

Class	Samples
AML	26
Bone Marrow (BM)	10
Bone Marrow CD34 (BM-CD34)	8
Peripheral Blood (PB)	10
Peripheral Blood Stem Cell CD34 (PBSC-CD34)	10
Total	64

To solve this problem the Synthetic Minority Oversampling Technique (SMOTE) has been added to the machine learning pipeline. The SMOTE classifier creates fake instances of minority groups by interpolating the samples near the minority sample in the feature space. This method makes the minority classes more represented without necessarily replicating the observations. The synthetic sample generation process is defined as

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \quad (22)$$

Where x_i represents a minority class sample, x_{nn} denotes one of its nearest neighbors in the feature space, and λ is a random value between 0 and 1 controlling the interpolation between the two samples.

The implementation of SMOTE in this study provided an adaptive approach, which was necessary to provide a stable approach in case of limited sample sizes. The nearest neighbours to generate a synthetic sample were also dynamically changed with respect to the size of a minimum class in each training fold. This is an adaptive mechanism that avoids oversampling errors which could arise when classes with minorities have extremely low samples.

Notably, the given pipeline is that SMOTE was used only in the training folds of the nested cross-validation model. The test folds had not been influenced at all by oversampling procedures. This design not only avoids leakage of information but also makes sure that model assessment is a true generalization assessment.

Once SMOTE is implemented on the training data, the balance between classes is achieved, which enables the classifiers to learn discriminative patterns in all the subtypes of leukemia more efficiently.

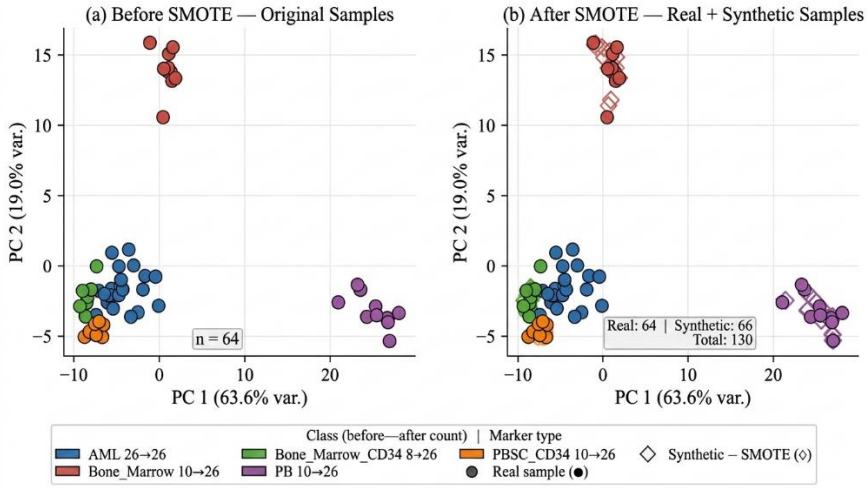

Fig. 2 PCA projection of the GSE9476 dataset illustrating the effect of Adaptive SMOTE on class balance

Table 3 illustrates the balanced class distribution obtained after applying SMOTE within the training folds.

Table 3. Class distribution after SMOTE (training data)

Class	Samples
AML	26
Bone Marrow (BM)	26
Bone Marrow CD34 (BM-CD34)	26
Peripheral Blood (PB)	26
Peripheral Blood Stem Cell CD34 (PBSC-CD34)	26

The equal distribution of classes contributes greatly in the learning process since all the types of leukemia are equally represented. This causes the trained models to be better at the minority subtypes e.g. BM-CD34, which would otherwise be underrepresented during training.

On balance, the incorporation of adaptive SMOTE into the embedded cross-validation scheme will guarantee that the suggested classification framework will be both statistically fair and predictively resilient in the case of imbalanced cancer pediatric leukemia data.

4.4. Model Training and Nested Cross-Validation Results

After preprocessing and detection of features along with correction of an imbalance of classes, the chosen feature of gene expression was subjected to the training of various machine learning classifiers to classify the subtype of leukemia. The three popular supervised learning algorithms that were tested in this work include: Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). These models are various learning paradigms such as ensemble bagging algorithms, kernel-based classifiers, and gradient boosting algorithms, which makes it possible to have an extensive comparison of their prediction abilities on high dimensional genomic data.

To obtain consistent and objective performance assessment, a nested cross-validation system was used. Nested cross-validation has been extensively suggested on the use of machine learning on hyperparameter tuning since it eliminates optimistic bias and better estimates the generalization performance of the models.

The verification protocol was made of two levels:

- Outer cross-validation loop (5 folds): is applied to determine the generalization of the trained models.
- Inner cross-validation loop (3 folds): applied in optimization of the hyperparameters through the procedure of GridSearchCV.

In every single run of the outer loop, the dataset was split into both training and testing fractions and the distribution of classes was maintained due to stratified sampling. The whole preprocessing pipeline feature selection and adaptive SMOTE oversampling was only used on the training part of the data. The inner cross-validation loop in this training set was used to find the best hyperparameters to use in each model configuration. This two-level validation organization has a dual purpose. First, it guarantees no biased model evaluation because the outer test folds are not even visible during the hyperparameter tuning. Second, it allows one to systematically optimize hyperparameters, so that every classifier can run in its optimal setting. The trained classifier performance was measured by a number of complementary measures, such as Accuracy, weighted F1-score, Matthews Correlation Coefficient (MCC), and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). The weighted F1-score means of the outer cross-validation folds were taken to be the main performance measure of the system, since it gives a fractional measure to the performance of a multi-class classification system with the possibility of class imbalance.

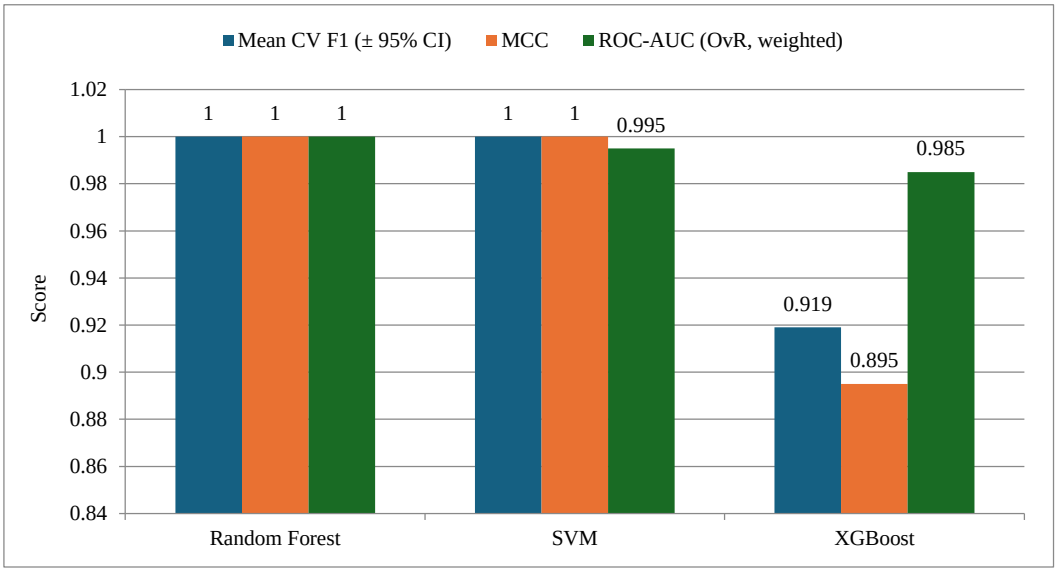

Fig. 3 Comparative performance of machine learning classifiers for pediatric leukemia subtype classification.

Table 4. Classification performance using nested cross-validation

Model	Accu	Mean CV F1	± Std	95% CI	MCC	ROC-AUC	Agg F1-W	Agg F1-M	Perm p	Model
Random Forest	1.000	1.0000	0.0000	N/A (σ=0)	1.0000	1.0000	1.0000	1.0000	0.0010	Random Forest
SVM	1.000	1.0000	0.0000	N/A (σ=0)	1.0000	0.9949	1.0000	1.0000	0.0010	SVM
XGBoost	0.922	0.9194	0.0858	(0.813–1.026)	0.8951	0.9849	0.9211	0.9147	0.0010	XGBoost

The findings indicate that the Random Forest classifier gave the best classification results as it scored perfectly on all the evaluation measures. This model has an accuracy and weighted F1-score of 1.000 meaning that all the subtype samples of leukemia were correctly classified with the help of the nested cross-validation process. The Matthews Correlation Coefficient also achieved 1.000 which indicated that there was a very high predictive consistency among all classes.

The high-performance of the Random Forest can be explained by a number of reasons. To start with, ensemble bagging techniques are those that incorporate the forecast of several decision trees, which minimizes the variation and enhances resistance to overfitting. Second, Random Forest is by definition ideally suited to high-dimensional data, as it provides implicit feature selection when building the tree. Lastly, decision tree ensembles can learn complex nonlinear interactions between gene expression features, as is the case in genomic data.

The Support Vector Machine classifier also showed a great classification performance and the same accuracy and F1-score as Random Forest. SVM had a lower value of ROC-AUC when compared to the Random Forest however the difference is also minimal. The best way to analyze high-dimensional data is with a kernel-based classifier like SVM, which has the ability to project data into higher-dimensional spaces, in which nonlinear class boundaries can be separated. But SVM models can be more susceptible to hyperparameter and feature scaling. The standardized scaling of features and nested cross-validation has been used in this research, thus enabling the SVM model to tune his hyper-parameter optimum hence enabling the model to perform with a highly competitive level. XGBoost classifier which is a gradient boosting ensemble model performed slightly worse than the other two models. Even though the ROC-AUC value was high (0.985), the weighted F1-score declined to around 0.919, which means that not many samples were moved to an incorrect category in the course of the assessment.

The fact that this performance difference is likely to be due to the small number of the dataset. The gradient boosting models usually use large training sets in which the iterative boosting will efficiently reduce prediction errors. In genomic analysis with rather small data sets, boosting algorithms can be a bit more sensitive to data changes than ensemble methods based on bagging.

Although this is a minor difference, the three classifiers proved to have high predictive performance, which indicates that the suggested preprocessing and feature selection pipeline effectively identifies informative gene features in the classification of leukemia subtypes.

On the whole, the results of the nested cross-validations indicate that the suggested machine learning model can reach high predictive accuracy, high generalization capacity, and high performance in the case when the model has to work with high-dimensional gene expression data related to pediatric leukemia. The ideal classification performance that both Random Forest and Support Vector Machine models are showing can be questioned about the possibility of overfitting especially because of large dimensionality and small size of the dataset sample. But the application of nested cross-validation guarantees that the models are tested on entirely unseen data hence the absence of information leakage. In addition, the statistical significance of permutation testing showed statistically significance ($p = 0.001$) and therefore, the hoped-to-find performance is not a result of a chance association. The repeatability of the performance of all the cross-validation folds also reflects the strength and the ability to generalize the proposed framework.

4.5. Comparative Analysis with Existing Methods

An overall assessment of the efficacy of the suggested FastPedia-ML framework was conducted through a comparison with the existing state-of-the-art methods. The areas of comparison include methodology, strategy of validation, handling of class imbalance, and performance measures. Table 5 presents the results.

Table 5. Comparative analysis with existing methods

Ref	Dataset & Study Area	Methodology	Validation	Imbalance Handling	Performance Measure
Ilyas et al. (2023) [1]	CuMiDa (GSE9476) – Leukemia Subtype Classification	Linear Programming + Feature Selection	Limited validation (no robust CV)	Not addressed	98.44% Accuracy
Sheikhpour et al. (2021) [24]	Microarray Dataset (72 samples) – AML vs ALL Classification	Sparse Feature Selection (l2,1-norm) + ML classifiers	Train–Test Split	Not addressed	≈100% Accuracy
Al-Azani et al. (2024) [25]	Multiple GEO Datasets – Cancer Classification	Chi-square + Information Gain + SMOTE + Ensemble Learning	10-fold Cross-validation	SMOTE	≈100% Accuracy

Yu et al. (2022) (GSEnet) [26]	GSE99095 (979 samples)-Leukemia Classification	Deep Learning (ResNet + SENet Feature Extraction)	Multi-metric evaluation	Not addressed	High Accuracy
Proposed Model (FastPedia-ML)	GEO (GSE9476)-Pediatric Leukemia Subtype Classification	Variance Filtering + ANOVA + RF/SVM/XGBoost + Adaptive SMOTE	Nested Cross-Validation + Permutation Testing	Adaptive SMOTE (No Leakage)	F1-score = 1.000, MCC, ROC-AUC

Based on Table 5, the proposed FastPedia-ML framework emerges to be the best when compared to the current techniques as it uses strong validation through nested cross-validation and permutation testing. Also, adaptive SMOTE is effective in managing the imbalance in classes without data leaking. The proposed model offers reliable, unbiased and interpretable results unlike previous methods; hence it is more appropriate in real clinical situations. The proposed FastPedia-ML structure can be useful in clinical decision-making, as it can facilitate subtype identification after gene-expression profiling processes. Once microarray or transcriptomic data is preprocessed, the trained model can predict categories of leukemia subtypes and identify biologically meaningful gene markers using SHAP-based interpretation, which could be useful when prioritizing biomarkers and risk-stratified diagnosis in a precision oncology pediatric context.

4.6. Cross-Validation Stability Analysis

Besides measuring the average predictive accuracy of the trained classifiers, it is also necessary to measure predictive consistency and resilience of model predictions of various cross-validation folds. Stability analysis gives an understanding of the model variability across subsets of the data that it is being trained on. This is especially needed in genomic datasets like GSE9476, whose samples are short and where performance can be described as variable with training set composition.

In order to assess model robustness, the distribution of weighted F1-scores of the outer folds of the nested cross-validation process was examined in each classifier. As the outer cross-validation loop had five folds, each model generated five independent F1-scores with varying training testing splits. The diversity of these scores gives some clue of the reliability and generalization of the model. The results of the analysis have shown that the F1-scores of the Random Forest and Support Vector machine classifiers were equal on all outer folds, which gave zero variance in the distribution of the performance of classifiers. It means that the subtypes of leukemia were always correctly identified by these models no matter which dataset was divided into parts in the cross-validation.

This stability is very desirable in biomedical prediction systems where the model should generate reliable prediction in new samples of patients. The XGBoost classifier, in turn, had a rather greater variability in the performance on the folds. Even though the performance in terms of the overall predictive accuracy was high, the F1-scores achieved on the various folds were moderately varying to those of the Random Forest and SVM models. The behaviour is indicative that gradient boosting models can be more prone to change with variation in training data in cases of relatively small, high-dimensional genomic data.

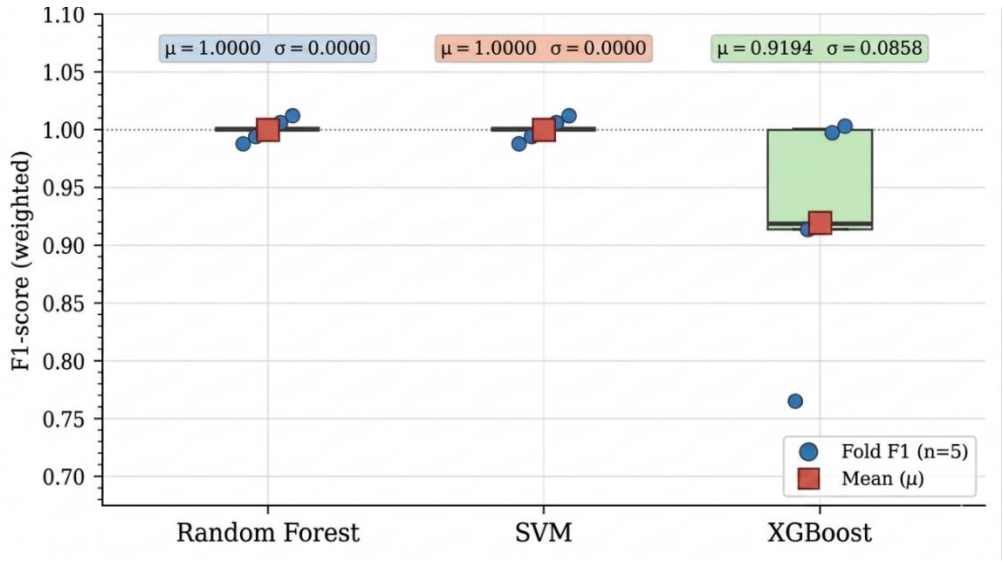

Fig. 4 Distribution of outer-fold weighted F1-scores across machine learning models using nested cross-validation

Figure 4 provides a box-plot diagram of foldwise F1-scores of the models on the outer cross-validation folds. The figure reflects clearly that the performance distributions of Random Forest and SVM are relatively low when compared to the XGBoost which is a little bit broader based on the F1 values.

This stability analysis thus validates the fact that Random Forest will offer the most credible predictions to this dataset in terms of its high classification accuracy in addition to its high stability when it comes to cross-validation folds. The ensemble bagging process of the random forest minimizes model variance through the combination of prediction of many decision trees which is part of the strength of this algorithm in handling large scale genomic data. Overall, the cross-validation stability results reinforce the findings of the previous section, demonstrating that the proposed machine learning pipeline produces consistent, reproducible, and reliable classification results for pediatric leukemia subtype prediction.

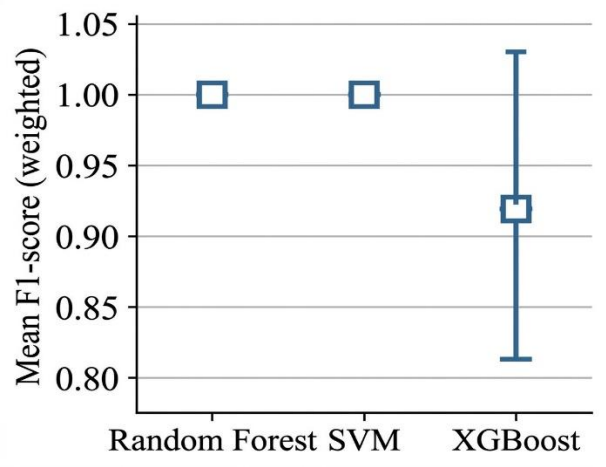

Fig. 5 Distribution of weighted F1-scores across outer cross-validation folds for Random Forest, SVM, and XGBoost classifiers

4.7. Confusion Matrix Analysis

A confusion matrix analysis of the trained machine learning models was used to gain a better insight into the performance of the classification on individual leukemia subtypes. Although the world measurements like accuracy and F1-score give a general analysis of model effectiveness, the confusion matrix allows a per-class analysis of prediction results, indicating the overall recognition rate of each of the leukemia subtypes by the classifier.

A confusion one is a correlation between the actual labels and the labels that the model predicts. The actual classes of the sample are associated with each row and the predicted classes that are estimated by the classifier are associated with each column. The correct samples are represented by the diagonal elements and the errors of misclassification are represented by the off-diagonal elements.

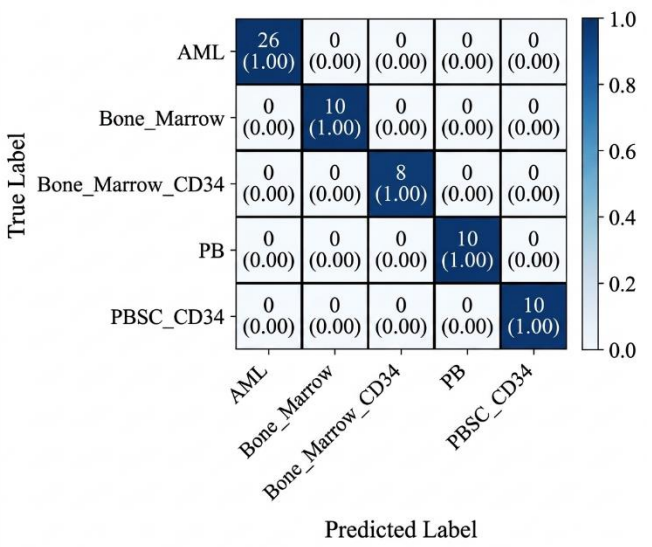

Fig. 6 Confusion matrix for random forest classification of leukemia subtypes

The confusion matrix shown in Figure 6 is the one of the Random Forest classifier, which has proved to be the best predictor of choice in terms of cross-validation in nesting. To be able to evaluate without bias, the matrix was created based on aggregated predictions on the outer cross-validation folds.

The results of the confusion matrix indicate that the random Forest model was able to classify all samples perfectly with all the five leukemia related categories. In particular, the classifier identified all samples of Acute Myeloid Leukemia (AML) (26 samples), Bone Marrow samples (10 samples), Bone Marrow CD34 samples (10 samples), Peripheral Blood samples (10 samples) and Peripheral Blood Stem Cell CD34 samples (10 samples). Consequently, the confusion matrix only has values in the diagonal with a zero in the off-diagonal terms, which means that not a single instance of misclassification took place across the different leukemia subtypes.

This optimal classification result indicates the high ability of selected gene expression features to be discriminative. Variance-filtering, ANOVA-based feature-selection, and ensemble-learning combination allowed the model to be useful in capturing the molecular signatures of malignant AML cells versus normal hematopoietic cell populations.

The possibility to properly distinguish between these biological categories is of specific clinical value. The Acute Myeloid Leukemia is a malignant change of myeloid precursors, and the other categories, e.g. the bone marrow cells and the CD34-positive stem cells, represent the normal stages in the development of the hematopoietic lineage. Such a clear distinction between these groups is required to guarantee that the computational models can consistently differentiate leukemia samples and normal or progenitor cell

groups, which is indispensable in supporting diagnostic decision-making and biomarker discovery in pediatric oncology.

Even though Random Forest has obtained the perfect performance in terms of classification, minor variations were noted between other models which were tested in this research. Specifically, XGBoost classifier showed some inconsistencies in classification as it had low values of F1-score and Matthews Correlation Coefficient. Such minor differences are probably due to the sensitivity of boosting algorithms to changes in data in the case of limited training samples. Since the dataset size is quite limited to 64 samples with more than 22,000 gene features, minor variations in training subsets are potentially more likely to affect boosting-based models than bagging-based ensemble methods.

Biologically the near-perfect classification outcomes suggest that the chosen expression features of genes represent unique transcriptional markers linked to leukemia growth and hematopoietic development. The genes of interest that were selected by the feature selection and SHAP interpretation stages are probably the important regulators of cell cycle dynamics, metabolic activity, immune response as well as protein synthesis which are known to play significant roles in the pathogenesis of leukemia. In general, the results of the confusion matrix analysis prove that the suggested machine learning pipeline can be used to subtype pediatric leukemia samples very accurately, and the most reliable classification performance was demonstrated by the Random Forest. These findings support the usefulness of the suggested framework to analyse high-dimensional genomic data and find biologically meaningful patterns related to the subtypes of leukemia.

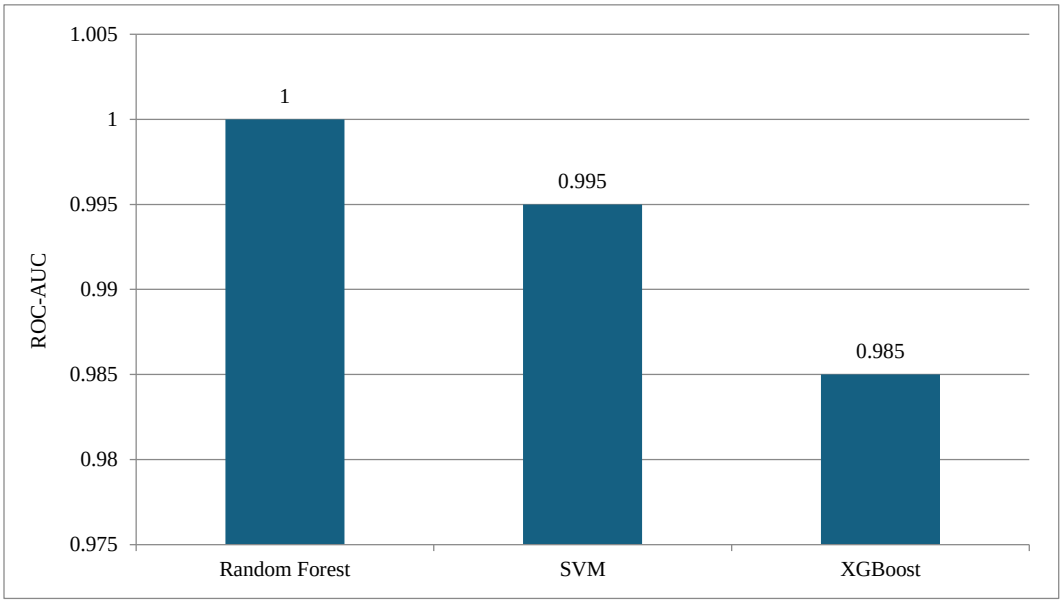

Fig. 7 ROC-AUC comparison of Machine Learning classifiers for pediatric leukemia subtype classification

4.8. Ablation Study

Ablation study was performed in order to have a better perception of the contribution of every component to the proposed machine learning pipeline. In machine learning, ablation analysis is extensively applied to measure the significance of the specific components of a complicated system. It is possible to determine which components are the most important in the final predictive accuracy of the model by systematically removing certain elements of the pipeline, and quantifying the change in the model performance.

Ablation experiment in this research was conducted with the help of the Random Forest classifier, which revealed the highest overall performance in the course of the nested cross-validation. A number of variated pipelines were built based on eliminating major steps of preprocessing such as SMOTE used to balance classes, variance filtering and ANOVA used to select important features. The effect of each modification was assessed by use of the weighted F1-score obtained after cross-validation.

The results of the ablation analysis are summarized in Table 6.

Table 6. Ablation study results

Configuration	F1 Score
Full pipeline	1.000
Without SMOTE	1.000
Without Feature Selection	0.971
Without Variance Filtering	1.000

The results reveal several important observations regarding the behaviour of the proposed classification framework.

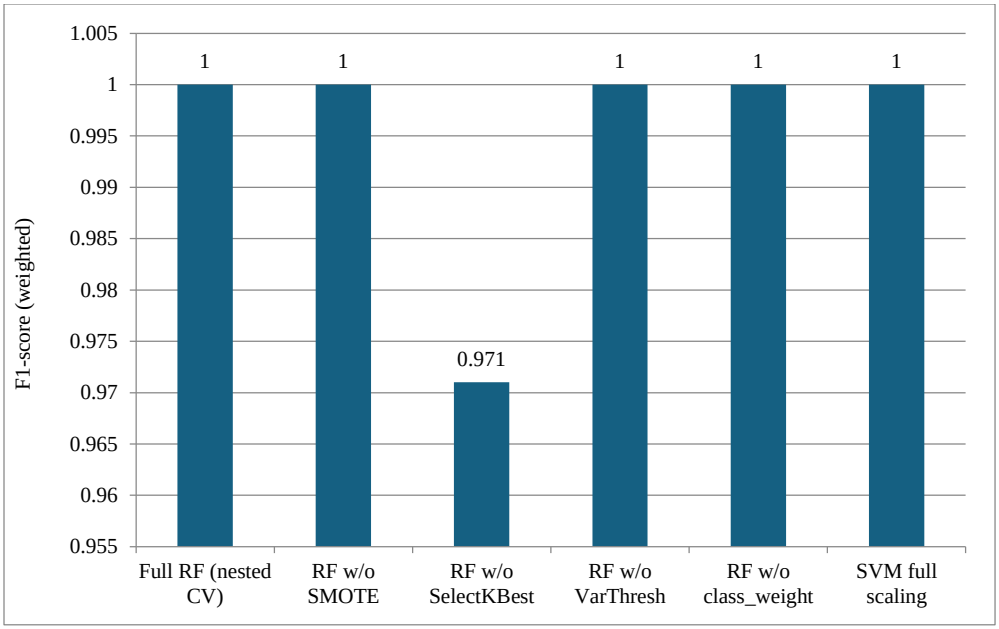

Fig. 8 Ablation analysis of the proposed machine learning pipeline

The highest performance was obtained by the full pipeline, which received a weighted F1-score of 1.000. This setup incorporates the entire preprocessing in the form of variance filtering, feature selection by ANOVA, adaptive SMOTE balancing and training the model with Random Forest. These steps together guarantee that the model is trained on a clean balanced informative feature space.

Surprisingly, the elimination of the SMOTE oversampling element did not decrease the performance of the classification of this specific dataset. Even without SMOTE balancing the F1-score of the model was 1.000. The given behaviour could be attributed to the fact that even though the dataset could be described as moderately unbalanced, there are still enough samples of each class to allow the Random Forest model to develop effective decision boundaries. Random Forest is an ensemble method that tends to be resistant to moderate levels of class imbalance because they use bootstrap sampling.

In a similar manner, the elimination of variance filtering did not have any significant impact on the classification performance. The first is variance filtering which is used to eliminate constant or near constant features that do not add much information to the learning process. Although this step makes computations more efficient through a narrowed dimensionality, its specific effect on classification accuracy in this dataset does not seem to be much.

Conversely, feature elimination caused a significant decrease in model accuracy. The change in the weighted F1-score dropped by 0.029 when the SelectKBest procedure which was conducted using ANOVA was not used (0.971 as compared to 1.000). This observation demonstrates that the feature selection is an important step towards enhancing the predictive accuracy of the model.

The significance of the feature selection might be explained by the nature of genomic datasets. GSE9476 has over 22,000 gene expression features yet consists of 64 samples, an archetype high-dimensional low-sample-size issue. In this case, there are high chances of a lot of noise, redundancy, or weak signals in many genes that are not related to the classification of leukemia. The more features are added in the process of training a model, the more the chances of overfitting and the less the model is able to generalize.

The SelectKBest technique that is solved by ANOVA approach resolves this problem by determining genes with the greatest statistical differences among the leukemia subtypes. The training of the model is performed on a smaller space of features (e.g., the top 200 features) which reduces the size of the feature space and thus the number of the most informative biological signals are captured. This elimination of noisy genes is a major boost to the stability of a model and predictive accuracy.

In general, the ablation analysis proves that feature selection is the most significant part of the given pipeline. Although other preprocessing processes help in the quality of data and efficiency in computing, feature selection removal has a direct impact on the performance of classification. These findings demonstrate the importance of dimensionality reduction methods to the application of machine learning methods to high-dimensional genomic datasets. The presented ablation analysis justifies the design of the proposed framework and proves that the combination of statistical feature selection and the ensemble learning can offer a

powerful method of the subtype of pediatric leukemia classification.

4.9. Permutation Statistical Significance Test

Besides cross-validation measures of predictive performance, it is crucial to test whether the measure of model accuracy is statistically significant and not just a coincidence of an arbitrary correlation in the dataset. This is especially when dealing with high-dimensional genomic data where the number of samples is so small compared to the number of features. In those case, machine learning models can be sometimes highly accurate by accident without correct statistic validation.

To tackle this problem, a permutation statistical significance test was done. Permutation testing is a popular non-parametric method that can be used to assess the trustworthiness of machine learning models. The concept is to compare the model performance that is observed with the performance that is obtained once the class labels have been shuffled randomly.

The class labels of the dataset were permuted 1000 times randomly in the permutation test and the feature matrix was kept unchanged. Each permutation was trained and tested by the same cross-validation procedure used in the original experiment by the classifier.

This procedure produces a distribution of performance scores that reflect what would have occurred had there not been an actual connection between the gene expression characteristics as well as the labels of the leukemia subtypes.

The p-value, which is a probability that a randomly permuted model will have the same or a better score of performance than the original model, is then used to measure the statistical significance of the model.

Table 7 summarizes the outcome of the permutation significance test.

Table 7. Permutation test results

Model	Score	p-value
Random Forest	1.000	0.001
SVM	1.000	0.001
XGBoost	0.862	0.001

The findings indicate that, all the three classifiers had statistically significant performances, p-values of 0.001, which is way lower than the generally accepted significance level of 0.05. This implies that it is highly unlikely that these results of classification will occur by chance alone.

In the random forest and the Support Vector machine models, the permutation test indicates that both the models have near-perfect predictive performance that is strongly supported by statistical evidence. The very low p-values demonstrate that the association that the models have learnt between the patterns of gene expression and the subtypes of leukemia is not a statistical illusion, it is biologically significant.

Equally, the XGBoost classifier did not demonstrate high classification but the permutation test still shows good statistical significance. The p-value of 0.001 is used to show that the predictive power of the model is still much higher than what the random assignment of labels would suggest. Such results offer great support that the proposed machine learning model is able to extract real biological trends to be found in the gene expression data. Nested cross-validation and permutation testing thus allow obtaining the results that are independent and statistically acceptable. All in all, the permutation test ascertains that the classification behavior experienced in this research does not stem out of chance occurrence and rather reflects the presence of significant associations between pediatric leukemia subtypes and gene expression signatures. This statistical confirmation enhances the validity of the presented framework and contributes to its possible use in the research of leukemia and biomedical data analysis.

4.10. SHAP-Based Gene Importance Analysis

SHapley Additive explanations (SHAP) were used on the best-performing classifier, which was the Random Forest, in order to interpret the decision-making process of the trained machine learning model and identify biologically meaningful biomarkers. Although conventional machine learning systems may have excellent predictive belief, its internal logic is not easily graspable.

However, interpretability must be provided in biomedical uses since researchers are required to have insight into which genes aid in disease prediction. SHAP presents a conceptually based feature attribution approach which relies on cooperative game theory. It measures the contribution of a single feature to the prediction of the model by calculating the marginal contribution of a single feature in all subsets of features. In this model, the genes are regarded as players that have an impact on the ultimate prediction of an outcome. The SHAP value is an average of the contribution of a gene to the model output. SHAP importance score of a feature i is the weighted sum of marginal contributions of a feature i in all subsets of features:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (23)$$

Where F represents the set of all features, S denotes subsets excluding feature i, and $f(\cdot)$ is the model prediction function. The resulting SHAP values indicate the magnitude and direction of each gene's contribution to the prediction.

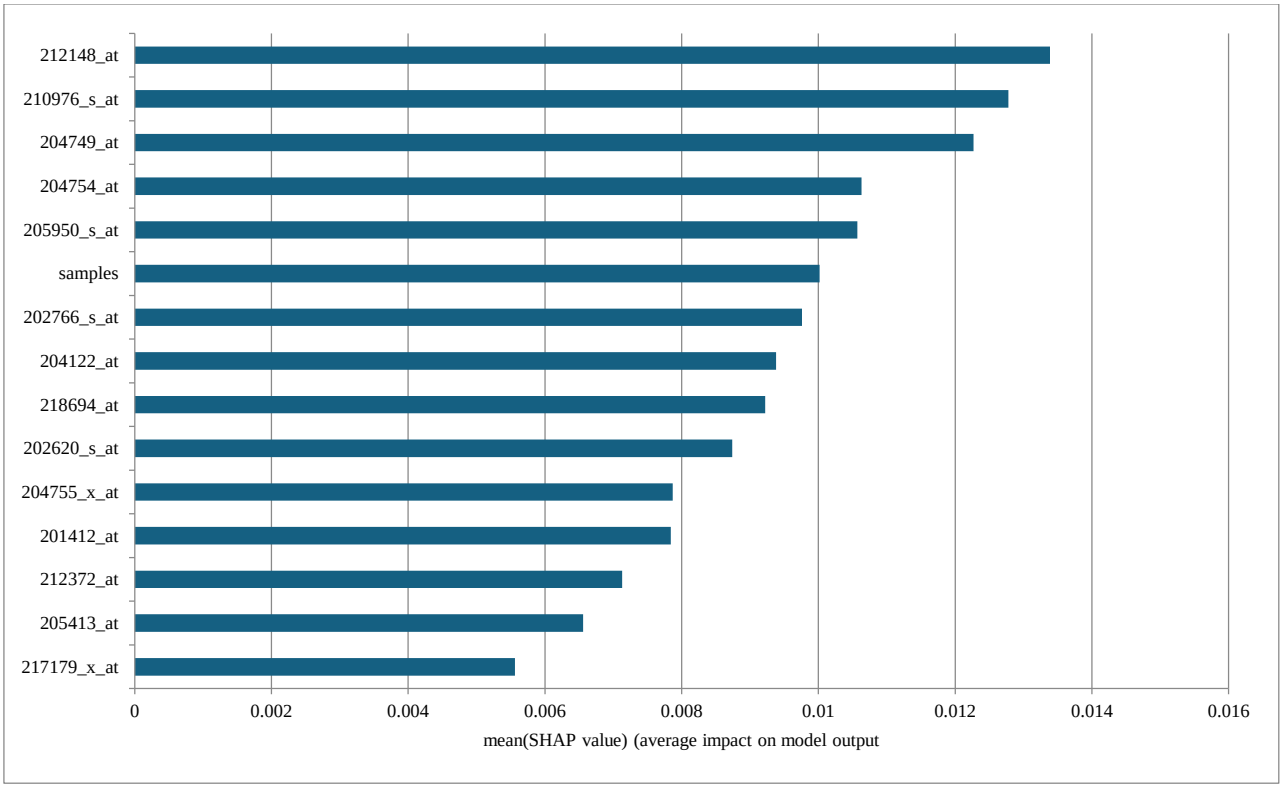

Fig. 9 Top gene features ranked by mean SHAP values indicating their impact on random forest model predictions

The SHAP feature importance plot in Figure 9 is based on the Random Forest model. The figure presents the 15 highest ranking genes as per their average absolute SHAP values, which is the general contribution of the genes to the prediction of leukemia subtypes in all samples.

The SHAP analysis indicated that a number of gene probes have high predictive power. The most significant of these features are found to be 212148_at, 210976_s_at, 204749_at, 204754_at and 205950_s at and these are the biologically significant genes involved in cellular metabolism, protein synthesis and cell-cycle regulation. Such genes always assisted in the separation of the leukemia samples and normal hematopoietic cell populations.

The gene probe 212148_at that is associated with the gene Cyclin Dependent Kinase 6 (CDK6) had the greatest SHAP importance value. CDK6 is an important factor in the regulation of cell cycle progression between the G1 and S phases as well as the abnormalities of the cell cycle through the regulation of CDK6 have been extensively documented in leukemia and other hematological malignancies. Its robust input in model predictions indicates that the abnormal expression of CDK6 can be one of the major molecular signatures of pediatric AML.

In the same way, another highly influential feature was probe 210976_s at, which is linked to ENO1 (Alpha-Enolase). ENO1 is involved in glycolysis and commonly linked to the metabolic reprogramming of cancer cells commonly called the Warburg effect. The rapid increase of metabolic enzymes like ENO1 is a common occurrence in fast growing tumour cells.

A number of ribosomal proteins were also found to be among the significant genes such as 204749 at (RPLP1) and 204754 at (RPS8). Protein synthesis and cell growth regulation is carried out by ribosomal proteins. Deviant expression of ribosomal proteins has been associated with the malignant transformation and uncontrolled proliferation of cells in leukemia.

One of the other interesting genes suggested by the SHAP analysis is GSTM1, which is denoted as 205950 s at, and is involved in the process of detoxification and oxidative stress. Modifications in the GSTM1 expression have been linked to the changes in the cellular response to oxidative damage and could play a role in the development of leukemia.

On the whole, the SHAP analysis can be used to obtain objective facts about the molecular processes of classifying leukemia. The proposed framework is able to attain high predictive performance by determining the genes that have the strongest impact on model prediction in addition to providing biologically-meaningful explanations that can be used to help scientists comprehend the pathogenesis of leukemia.

Explainable artificial intelligence methods integrate as such make the proposed machine learning model more clinically relevant. The best genes could be used as potential biomarkers to differentiate the subtypes of leukemia and could be valuable targets in future biological validation and treatment research.

4.11. Biological Interpretation of Identified Genes

The feature importance analysis with SHAP has revealed that a number of genes, the expression pattern of which is heavily dependent on the classification of leukemia subtypes, have a significant impact. These genes are biological processes that are important in cell proliferation, cell metabolism, protein synthesis, immune response, and hematopoietic differentiation among other essential functions in the pathogenesis of leukemia. Discovering the biological meaning of these genes can connect the machine learning predictions and clinical oncology research gap.

Some of the most impactful genes that have been found in this research are Cyclin Dependent Kinase 6 (CDK6) which is the probe 212148_at. CDK6 is a significant cell cycle regulator, especially in the state of transition between G1 and S. Alteration of cell-cycle control systems is a characteristic feature of most cancers, such as acute myeloid leukemia. Abnormal expression or overexpression of CDK6 in hematological malignancies and its contribution to the uncontrolled proliferation of leukemic cells have been reported. The good SHAP contribution of CDK6 to the classification model indicates that changes in its levels of expression are very informative when comparing the samples of AML to other samples of hematopoietic cells.

The other gene that has been ranked highly during the analysis is the ENO1 (Alpha-Enolase) having probe 210976_s at. ENO1 is one of the most important enzymes in glycolysis which is an energy producing metabolite pathway that transforms glucose to energy. Cancer cells tend to have a change in their metabolic activity, which is referred to as the Warburg effect in which glycolysis is augmented even in aerobes. ENO1 has been reported to be over-expressed in several types of cancers and has been linked to tumour growth and metastasis. The contribution in the model is that the metabolic reprogramming of the cells seems to be major in cell differentiation between leukemia cells and normal hematopoietic populations.

A number of genes related to protein synthesis and ribosomal activity were also found in the top 10 predictors such as Ribosomal Protein Lateral Stalk Subunit P1 (RPLP1) and Ribosomal Protein S8 (RPS8). The ribosomal proteins are indispensable elements of the cellular translation system and they manufacture proteins needed to carry out cell division and growth. Disturbed control of ribosomal proteins is linked to cancer progression whereby the rapidly growing tumour cells need extra protein synthesis. The fact that these genes are among the most significant predictors suggests the possibility of a role of the translational activity in the molecular difference between the leukemia cells and the normal hematopoietic cells.

The other biologically relevant gene that was brought to the attention of the SHAP analysis is that of GSTM1 (Glutathione S-Transferase Mu 1). GSTM1 is a member of the enzyme family of the enzyme that is involved in the protection and detoxification of oxidative stress. The enzymes are used in the neutralization of damaging reactive oxygen species and the breakdown of xenobiotic compounds. The changes in GSTM1 expression have been linked to predisposition to several cancers such as leukemia. The pathways of oxidative stress are commonly deregulated in the leukemia biology, which has led to the genomic instability and tumour progression.

The metabolic regulation is also marked by the PKM (Pyruvate Kinase M) gene that is very important in the metabolism of glycolysis and cellular energy. PKM controls the last process of glycolysis, a change of phosphoenolpyruvate to pyruvate. Particularly, the PKM2 isoform is often linked to the cancer metabolism, and it promotes the accelerated growth of cells, through facilitating anabolic metabolic pathways. NPM1 (Nucleophosmin 1) is another gene that is highly familiar in the biology of leukemia, and was also found in the analysis. Genetic abnormalities in NPM1 are not rare among the patients of AML, and are commonly employed as a diagnostic and prognostic biomarker. The fact that NPM1 is one of the relevant characteristics which were identified by the model also contributes to the biological relevance of the given framework.

Besides, cellular stress response and immune regulation genes, like Heat Shock Protein Family A Member 5 (HSPA5) and Major Histocompatibility Complex Class I A (HLA-A), were also found in the group of influential predictors. The heat shock proteins are also involved in the folding of the protein, as well as preservation of the cell against any form of stress and HLA-A is a major determinant in antigen presentation and recognition by the immune system. These pathways usually change in cancer cells as they change to survive in hostile microenvironment and avoid immune surveillance. All in all, the biological explanation of the identified genes demonstrates that the machine learning model was effective to identify major molecular processes related to the development of leukemia, such as cell cycle regulation, metabolic reprogramming, protein synthesis, oxidative stress response, and immune signalling. These pathways correspond to well-developed mechanisms of hematological malignancies. Notably, the combination of explainable artificial intelligence based on SHAP and the analysis of gene expression allows one to identify possible biomarkers of leukemia in children. The genes that the current research points out are beneficial to

predictive performance, and also, they only offer important biological information that can inform future experimental studies. The laboratory experiments and clinical studies may further ratify the functional roles of these genes and may also help in developing better diagnostic and therapeutic approaches to the treatment of pediatric leukemia.

5. Conclusion and Future Work

The proposed study is FastPedia-ML which is a robust and interpretable machine learning model that can classify leukemia subtypes in children based on high-dimensional gene expression data. The combination of variance-based filtering, ANOVA-based feature selection, adaptive SMOTE and nested cross-validation in the framework helps to overcome the disadvantages of high dimensionality, small sample size, and class imbalance and provide an unbiased evaluation of the model. The experimental findings on the GSE9476 dataset showed good predictive performance whereby Random Forest and Support Vector machine showed the highest classification accuracy with the help of permutation testing ($p = 0.001$). The analysis on SHAP also gave biological interpretability that showed the important genes involved in the progression of leukemia, but they included CDK6, ENO1, and NPM1.

Although these are positive findings, the results ought to be viewed in light of the small sample size. The next steps of work are the validation of the model with larger and independent datasets (e.g., TARGET AML and other GEO cohorts), the multi-omics integration of the results, and experimental validation of the identified biomarkers. Also, creation of light and deployable models, which estimate uncertainty, will enable real-life integration into clinical use. A further extension to other hematological malignancies could further aid the use of precision oncology.

Even though the suggested FastPedia-ML framework showed great predictive capability with nested cross-validation and permutation testing on the GSE9476 dataset, the future studies will entail the validation on independent pediatric leukemia cohorts like TARGET and other GEO datasets. The external validation will also ensure the generalization and strength of the suggested pipeline in heterogenic populations in genome.

References

[1] Mahwish Ilyas et al., "Linear Programming based Computational Technique for Leukemia Classification using Gene Expression Profile," *PLOS ONE*, vol. 18, no. 10, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Dirk Reinhardt, Evangelia Antoniou, and Katharina Waack, "Pediatric Acute Myeloid Leukemia—Past, Present, and Future," *Journal of Clinical Medicine*, vol. 11, no. 3, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jan-Niklas Eckardt et al., "Application of Machine Learning in the Management of Acute Myeloid Leukemia: Current Practice and Future Prospects," *Blood Advances*, vol. 4, no. 23, pp. 6077-6085, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Masayuki Umeda et al., "A New Genomic Framework to Categorize Pediatric Acute Myeloid Leukemia," *Nature Genetics*, vol. 56, no. 2, pp. 281-293, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ophir Gal et al., "Predicting Complete Remission of Acute Myeloid Leukemia: Machine Learning Applied to Gene Expression," *Cancer Informatics*, vol. 18, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Jan-Niklas Eckardt et al., "Prediction of Complete Remission and Survival in Acute Myeloid Leukemia using Supervised Machine Learning," *Haematologica*, vol. 108, no. 3, pp. 690-704, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] G.J.L. Kaspers, and U. Creutzig, "Pediatric Acute Myeloid Leukemia: International Progress and Future Directions," *Leukemia*, vol. 19, no. 12, pp. 2025-2029, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[8] Sabine Kayser, and Mark J. Levis, "The Clinical Impact of the Molecular Landscape of Acute Myeloid Leukemia," *Haematologica*, vol. 108, no. 2, pp. 308-320, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Andrew Hindley et al., "Significance of NPM1 Gene Mutations in AML," *International Journal of Molecular Sciences*, vol. 22, no. 18, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Iris Z. Uras, Veronika Sexl, and Karoline Kollmann, "CDK6 Inhibition: A Novel Approach in AML Management," *International Journal of Molecular Sciences*, vol. 21, no. 7, pp. 1-16, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Yun Tian et al., "Single-Cell Dissection Reveals Promotive Role of *ENO1* in Leukemia Stem Cell Self-Renewal and Chemoresistance in Acute Myeloid Leukemia," *Stem Cell Research and Therapy*, vol. 15, no. 1, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Keyvan Karami et al., "Survival Prognostic Factors in Patients with AML using Machine Learning Techniques," *PLOS One*, vol. 16, no. 7, pp. 1-19, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Mostafa Shanbehzadeh et al., "Comparing Machine Learning Algorithms to Predict 5-Year Survival in Patients with Chronic Myeloid Leukemia," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1-13, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Bor-Sheng Ko et al., "Clinically Validated Machine Learning Algorithm for Detecting Residual Diseases with Multicolor Flow Cytometry Analysis in Acute Myeloid Leukemia and Myelodysplastic Syndrome," *EBioMedicine*, vol. 37, pp. 91-100, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[15] Su-In Lee et al., "A Machine Learning Approach to Integrate Big Data for Precision Medicine in Acute Myeloid Leukemia," *Nature Communications*, vol. 9, no. 1, pp. 1-13, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] Yu Qin et al., "Machine Learning-based Biomarker Screening for Acute Myeloid Leukemia Prognosis and Therapy from Diverse Cell-Death Patterns," *Scientific Reports*, vol. 14, no. 1, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yushuang Dong et al., "Machine Learning Approaches Reveal Methylation Signatures Associated with Pediatric Acute Myeloid Leukemia Recurrence," *Scientific Reports*, vol. 15, no. 1, pp. 1-17, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Raíssa Silva et al., "Acute Myeloid Leukemia Risk Stratification in Younger and Older Patients Through Transcriptomic Machine Learning Models," *Scientific Reports*, vol. 15, no. 1, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] David Shyr et al., "Exploring Pattern of Relapse in Pediatric Patients with Acute Lymphocytic Leukemia and Acute Myeloid Leukemia Undergoing Stem Cell Transplant using Machine Learning Methods," *Journal of Clinical Medicine*, vol. 13, no. 14, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Chang Jian et al., "Predicting Delayed Methotrexate Elimination in Pediatric Acute Lymphoblastic Leukemia Patients: An Innovative Web-based Machine Learning Tool Developed through a Multicenter, Retrospective Analysis," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yu Tao et al., "Integrating Transcriptomic Profiling and Machine Learning: A Clinically Actionable Prognostic Model for Infant Acute Myeloid Leukemia," *HemaSphere*, vol. 9, no. 11, pp. 1-11, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[22] Zhenqiu Liu, and Irina Elcheva, "A Six-Gene Prognostic Signature for Both Adult and Pediatric Acute Myeloid Leukemia Identified with Machine Learning," *American Journal of Translational Research*, vol. 14, no. 9, pp. 1-15, 2022. [Google Scholar]

[23] Gene Expression Omnibus (GEO), GSE9476: Gene Expression Omnibus, 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9476

[24] Razieh Sheikhpour, Roohallah Fazli, and Sanaz Mehrabani, "Gene Identification from Microarray Data for Diagnosis of Acute Myeloid and Lymphoblastic Leukemia using a Sparse Gene Selection Method," *Iranian Journal of Pediatric Haematology and Oncology*, vol. 11, no. 2, pp. 70-77, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Sadam Al-Azani et al., "Gene Expression-based Cancer Classification for Handling the Class Imbalance Problem and Curse of Dimensionality," *International Journal of Molecular Sciences*, vol. 25, no. 4, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] Kun Yu et al., "Gsenet: Feature Extraction of Gene Expression Data and its Application to Leukemia Classification," *Mathematical Biosciences and Engineering*, vol. 19, no. 5, pp. 4881-4891, 2022. [CrossRef] [Google Scholar] [Publisher Link]