Enhanced Leukemia Subtype Classification Using SMOTE and Hybrid Feature Selection in Microarray Data

Chaitra P C, R Saravana Kumar
Dayananda Sagar Academy of Technology and Management, Visvesvaraya Technological University, Belagavi, 590018
E-mail: chaitrapcjay@gmail.com, saravanakumar.rsk28@gmail.com

Keywords: leukemia, multi class classification, class imbalance, feature selection

Received: September 18, 2025

Blood cancer is a rising issue in the past decade, and early detection is a must for early intervention. Traditional techniques for diagnosing blood cancer include high expense, long processes, and medical professionals and a variety of tests. Hence, an effective prediction model with high accuracy is a must. This study presents a robust leukemia multiclass classification framework leveraging advanced ML (machine learning) techniques. Addressing key challenges such as class imbalance, high-dimensional gene expression data, and feature selection. This study presents an integrated approach for data balancing by combining the Synthetic Minority Oversampling Technique (SMOTE) with nonlinear interpolation. A hybrid feature selection model utilizing Principal Component Analysis (PCA) on Linear Discriminant Analysis (LDA) is implemented to enhance classification performance. Experimental results indicate that using SMOTE with PCA+LDA on Random Forest classifiers outperforms traditional methods, achieving 98% accuracy in leukemia multiclass classification.

Povzetek: Študija predstavlja učinkovit model strojnega učenja za večrazredno razvrščanje levkemije, ki z uporabo SMOTE, PCA+LDA in naključnega gozda doseže 98 % točnost.

1 Introduction

Cancer is defined as a collection of conditions that emerge when a cell's growth becomes uncontrolled, enabling it to spread to other locations throughout the body. WHO [20] reports that cancer remains one of the greatest health issues globally, contributing to almost 10 million deaths in the year 2020, accounting for approximately one out of six deaths around the globe. Early detection, because this substantially raises prospects for effective treatment alongside survival, underlines necessity for continuous checks and immediate medical attention. Continuous innovations in treatment and palliative care serve as motivation for decreasing the global cancer burden [1].

In that Leukemia is a malignant disease affecting white blood cells disrupts the normal functioning of the bone marrow. Leukemia is a condition that affects the body's capacity to operate normally by replacing healthy blood cells with aberrant cancer cells. A genetic mutation in an immature blood cell transforms it into a cancerous cell, which proliferates abnormally, outliving normal cells and rapidly multiplying.

Leukemia is categorized according to the type of white blood cells impacted and the rate of progression. It could be acute, getting worse quickly, or chronic, taking longer to develop. Furthermore, the afflicted cells classify it as either myeloid or lymphoid. Leukemia comes in four primary forms which include: The majority of patients with ALL (acute lymphoblastic leukemia) are youngsters. AML (Acute Myeloid Leukemia): A common condition in both children and adults. The majority of patients having CML (chronic myeloid leukemia) are adults. CLL (Chronic Lymphocytic Leukemia): Occurs mainly in older adults. Timely diagnosis with proper subtypes is critical for effective treatment. Early-stage treatments focus on disease management, while advanced-stage interventions aim to eradicate leukemic cells, allowing the bone marrow to resume normal blood cell production.

Clinical Motivation for Multiclass Leukemia Sub-Type Classification Leukemia diagnosis requires distinguishing among biologically distinct subtypes such as Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Myeloid Leukemia (CML), and Chronic Lymphocytic Leukemia (CLL), as well as finer microarray-defined subclasses (e.g., ALL-B, ALL-T, MLL). These subtypes differ significantly in prognosis, treatment protocols, and survival outcomes. Microarray-level subtypes often exhibit overlapping gene-expression signatures, making subtype separation a challenging multiclass task, especially when sample sizes are heavily imbalanced. A multiclass framework is therefore clinically essential, unlike bi- nary ALL–AML classifiers, because oncologists require accurate subtype identification before initiating therapy.

Bioinformatics has been comparatively recent bi- ology domain which integrates algebraic, analytical, alongside computational techniques for processing alongside interpreting biological data. Broadly defined, bioinformatics involves the use of digital technologies to analyze high-dimensional biological datasets collected from diverse sources. Advances in DNA microarrays and proteomics for large-scale gene expression studies have improved the technology, and with this, bioinformatics tools have become increasingly important. Modern

research combines laboratory experiments with bioinformatics analyses. The molecular profiling of cancer samples has become increasingly critical for both the advancement of cancer research and better treatment approaches.

In this Microarray technology, the laboratory technique studies biological molecules, especially DNA and proteins. The probes of interest are attached to a small glass, plastic, or silicon chip in a grid pattern. Microarrays allow the simultaneous analysis of thou- sands of phenomena within a sample, such as detecting specific genetic sequences, analyzing gene expression levels, and studying protein interactions. Microarray technology has greatly improved cancer classification. However, it has challenges, too, because it analyses small sample sizes, with each sample containing many genes. This challenge, known as the curse of dimensionality, is further compounded via large quantity of uninformative genes, which may actually reduce classification performance. Behaviors like these lead to common recommendations such as filtration and feature selection. These steps help ensure that only the most significant genes with differential expression be- tween outcome classes are used for classifier construction. This, in turn, motivates investigating different classification techniques that could be utilized for identifying various types and stages of leukemia and ultimately help with early detection and individual treatment to minimize mortality.

Paper structure has been given as follows: Along with research on several cancer classification algorithms, Section 2 addresses important issues like class imbalance, high dimensionality, and the requirement for efficient feature selection. A summary of the public datasets used for experimentation is given in Section 3, with an emphasis on class imbalance and high dimensionality in gene expression data to enhance the classification of leukemia subtypes (e.g., ALL, AML, CLL, CML). Model validation and experimental findings are shown in Section 4. Section 5 wraps up the analysis and suggests possible avenues for further investigation.

More text of the introduction. More text of the introduction. More text of the introduction. More text of the introduction.

2 Literature survey

Text of the second section. Classifying cancer utilizing microarray gene expression data has received growing attention in recent years. Current progress in ML alongside DL (deep learning) significantly improved classification accuracy. Epidemiological studies project a 12.8% increase in cancer cases in India by 2025, underscoring the need for improved computational techniques [1]. B. Mabrouk et al. explored PCA (Principal Component Analysis) as well as LDA utilisation for dimensionality reduction, emphasizing LDA's effectiveness in improving class separability for Alzheimer's disease classification [2]. P. K. Mallick et al. show that DL models for leukemia's binary classification, including DNNs (Deep Neural Networks) alongside CNNs (Convolutional Neural Networks), exhibited high accuracy in leukemia classification, with a five-layer DNN, achieving 98.2% accuracy in distinguishing ALL from AML [3]. Ensemble learning techniques, such as bagging with Multilayer Perceptrons (MLPs) and mutual information- based feature selection, have been applied to gene expression datasets, significantly enhancing classification performance [4] [5]. U. Ravindran and C proposed Hybrid approaches combining data augmentation methods like Wasserstein Tabular Generative Adversarial Networks (WT-GAN) with DL have improved cancer detection beyond 97% accuracy [6]. Integrated RNA-seq and microarray analysis have provided better leukemia classification, identifying significant gene signatures such as BLK, DOCK2, and RPS15 [6]. Feature selection approaches like Chi2, and SMOTE-Tomek have been employed to handle class imbalance and high-dimensional datasets, achieving remarkable classification results with 96.7% ac- curacy in leukemia detection [8]. Sadam AI-Azani et. Al proposed two ways of handling class imbalance moreover curse of dimensionality by data level and algorithm level. According to experiments on leukemia-ALLAML, leukemia subtype, colon tumour, along with leukemaia CuMiDa, SVM-SMOTE oversampling methodology in conjunction with random forests produced best outcomes with respect to other assessed oversampling alongside ensemble learning methods [9]. Abdul Karim et. Al performed experiment on GSE9476 by combining LR, SVM, DT using voting classifiers referred to as LDSVM with hard voting, attaining an accuracy of 0.95, and proposed that the algorithm tested in ML is capable of transforming the traditional classification methods for application in the medical field or, in general, bioinformatics [10]. Vaibhav Rupapara et al. addressed microarray challenges using ADASYN resampling as well as Chi2(Chi- squared) feature selection approaches. They proposed a hybrid classifier, Logistic Vector Trees (LVTrees), which integrates Logistic Regression, Support Vector Classifier, and Extra Trees Classifier. Their study demonstrated that LVTrees outperform traditional approaches. They further recommended exploring the integration of multiple datasets to develop a complex, high-dimensional dataset for enhanced evaluation of the proposed method [11]. In cancer classification, most of the algorithms are working well with binary classification and the need for more advanced techniques to deal with the multi-class classification of cancers. These advancements emphasize the critical role of ML and DL in enhancing cancer classification ac- curacy, optimizing computational efficiency, and integrating multi-source gene expression data for improved diagnostic capabilities.

Recent work in hybrid feature selection for cancer genomics includes mRMR-SVM pipelines, ReliefF with ensemble learners, sparse LDA variants, and imbalance-aware learning frameworks such as SMOTE-ENN, ADASYN-SVM, and GAN-based augmentation. Several studies between 2020–2025 demonstrate the effectiveness of combining oversampling with supervised feature selection for microarray datasets.

3 Materials and methods

This section explains the dataset used for the classification and methodology followed to predict the cancer with better accuracy.

3.1 DATA SET details

Dataset Description: The experiments use the pub- licly available leukemia microarray dataset GSE13159 from the Microarray Innovations in Leukemia (MILE) study. The dataset contains 2,096 samples profiled us- ing the Affymetrix HG-U133 Plus 2.0 platform with 54,675 probe sets. For this study, five diagnostic classes were included: ALL (n=750), AML (n=542), CLL (n=488), CML (n=76), and non-leukemia controls (n=280). Only this single dataset was used; no cross-platform merging was performed. All experiments used stratified 5-fold cross-validation to avoid biased splits.

Dataset is downloaded from the public repository from https://ncbi.nlm.nih.gov/ with Gene accession number GSE13159. Gene expression profiling using microarrays is used to diagnose and subclassify leukemia. This dataset is a component of the MILE (Microarray Innovations in Leukemia) research pro- gram, which is financed by Roche Molecular Systems and led by the European Leukemia Network (ELN).

For the Affymetrix HG-U133 Plus 2.0 GeneChips, 2096 blood or bone marrow samples from individuals with acute and chronic leukemia were hybridized. Five subtypes of leukemia cancer—ALL, AML, CML, and CLL—as well as non-leukemia instances are included in the data collection.

3.2 Methodology

The methodology proposing comprehensive pipeline of classification is explained in Figure 1. The procedure begins with high-dimensional gene expression matrices, in which samples are represented by columns and genes are represented by rows. Preprocessing steps include normalization to standardize data, handling missing values, and addressing class imbalance through effective data sampling. Feature engineering proceeds, including extracting the most critical genes and further dimensionality reduction in order to filter out unnecessary information. The processed dataset is divided into training (70%) and test data(30%), and the training data is used to build both traditional ML models. Finally, performance analysis evaluates the models using metrics like precision, accuracy, recall, and F1-score. This pipeline provides an end-to-end framework for leveraging gene expression data to en- hance leukemia diagnosis and classification, combining data preprocessing, feature engineering, and ML techniques to improve accuracy and robustness.

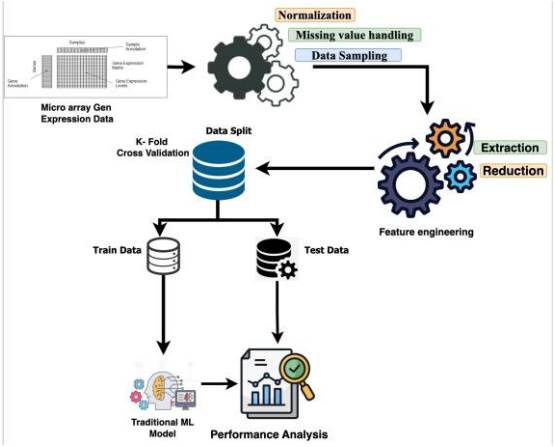

Figure 1: Methodology used for Leukemia cancer classification using microarray gene expression data

Text of the second section. Prevention of Data Leakage To ensure valid and reproducible results, all preprocessing steps—SMOTE oversampling, PCA, and LDA—were performed strictly within each training fold of the stratified cross- validation procedure. The test fold remained completely unseen during oversampling and feature extraction to prevent information leakage. This ensures that the reported results represent leakage-free generalization performance.

3.2.1 Imbalance data handling

Sample imbalance arises in a dataset when one of the groups contains more samples compared to another group. The challenge is much more significant in gene expression studies, mainly on datasets related to cancer. Sample imbalance can arise from several factors, such as the scarcity of available cancer samples, bud- get, and purposeful reduction of the control group samples. Since gene expression data are already affected by small sample sizes, the problem of sample imbalance can even worsen the analytical difficulties. From Literature survey found that various adaptations of the Synthetic Minority Over-Sampling Technique (SMOTE) have been introduced to address sample imbalances and observed that improved accuracy and robustness of differential gene expression analysis. While the standard SMOTE technique applies linear interpolation between minority instances, In this work nonlinear variations introduced further enhance the quality of generated samples. SMOTE selects a sample and its k nearest neighbors, then uses interpolation to produce additional data points to build synthetic samples for the minority class. In this reference, the algorithm of SMOTE over-sampling algorithm was developed by Chawla et al. (2002) [12].

Algorithm 1: SMOTE with Non-Linear Interpolation

Input: Imbalance Dataset X with n sample and features
Output: Dataset With more realistic synthetic sample added in minority Class

repeat

Choose a minority class sample x_0.
Identify the K-nearest neighbors of x_0 from the minority class.
Randomly pick one sample x_k from the identified K-nearest neighbors, where k represents its rank.
Apply a non-linear interpolation between x_0 and x_k to create a new synthetic data point z using polynomial function P(X) to generate new samples:

$$z = x0 + \lambda P (xk - x0) \quad (1)$$

Fit a polynomial function using x_0 and its neighbors.
Compute the interpolation using degree d using quadratic.
Generate synthetic data using the polynomial equation.

Until M synthetic samples are generated.

SMOTE Configuration For oversampling, SMOTE was configured with k=5 nearest neighbors and a sampling strategy that balances all minority classes to the size of the majority class (ALL). A quadratic (degree-2) interpolation function was used for nonlinear SMOTE. Borderline-SMOTE and ADASYN were also evaluated; however, classical SMOTE with nonlinear interpolation provided more stable minority-class boundaries and avoided generating noisy edge samples.

SMOTE non-linear interpolation can model com- plex decision boundaries. Avoids creating overlap- ping synthetic samples. Table 1 represents the sample count of each class before and after applying the sampling methods, and the result is highlighted in bold.

Table 1: Class distribution of leukemia subtypes before and after applying SMOTE.

Leukemia Subtype	Before SMOTE	After SMOTE
ALL	750	750
AML	542	542
CLL	488	488
CML	76	456
Others	280	280

Biological Consistency of Nonlinear SMOTE: To ensure that synthetic samples do not deviate from biologically plausible gene-expression patter. the nonlinear SMOTE mechanism is restricted to local neighbourhoods within the same leukemia subtype. The polynomial interpolation function is fitted using only k-nearest minority samples, ensuring that synthetic samples remain within the intrinsic gene- expression manifold of that class. This prevents the creation of unrealistic gene combinations and maintains relative expression order among biologically co- regulated genes. Such neighbourhood-preserving interpolation has been shown to approximate the natural nonlinear variability present in microarray datasets, thereby improving the biological validity of synthetic data.

3.2.2 Dimensionality reduction

Dimensionality reduction is a crucial analytical technique that reduces dataset features while retaining essential information. It enhances ML model efficiency, reduces overfitting, and improves data visualization. PCA and LDA are commonly used for this purpose.

Hybrid Model (PCA on LDA) In this proposal, a hybrid model works better than the individual performance in the gene expression data by gene extraction followed by gene selection. Steps to be followed: First, apply the PCA application to reduce the dimensions while maintaining 95 percent variance. Next, Principal Component Extraction, i.e., transformation of the original data points into a lower-dimensional space. Then, on the PCA outputs, apply LDA to employ the class labels in seeking axes that separate these classes. Linear Discriminant Component Extraction reduces dimensions further down to the number of classes. Finally, Classification is done on transformed. Clarification of PCA–LDA Order PCA was ap- plied first to reduce the original 54,675-dimensional gene-expression matrix into a compact representation retaining 95% of the variance. LDA was subsequently applied on the PCA-transformed features to maximize class separability using diagnostic labels. LDA reduces the dimensionality to C–1 = 4 components for the five leukemia classes. This order—PCA → LDA—is standard for microarray data because PCA mitigates noise and multicollinearity before supervised discriminant learning.

3.2.3 Principal component analysis (PCA)

PCA has been unsupervised approach for reducing dimensionality by projecting data onto a lower-dimensional space while preserving maximum variance. It is commonly used for preprocessing before ML. The core concept involves identifying principal components—new feature axes that capture the most variation in the dataset [14]. PCA follows a systematic mathematical approach involving several key steps. First, the standardization of data takes place. Since the various features in a dataset could be on different scales, PCA starts off by preprocessing this data. In this way, all features contribute equally to the analysis. The eq. 2 gives standardizing data i.e. mean subtraction from the data divided by standard deviation:

$$X_{standardized} = \frac{X-\mu}{\sigma} \quad (2)$$

Where μ denotes mean of every feature, σ denotes standard deviation. In the second step, Compute Co-variance Matrix. Covariance matrix denotes relation- ship amongst different features. It is given by eq. 3.

$$C = \frac{1}{n-1} (X^T X) \quad (3)$$

where X is the standardized dataset.

where X is the standardized dataset. Followed by computing Eigenvalues as well as Eigenvectors. Eigen- vectors denote maximum variance directions. How- ever, eigenvalues measure variance by all principal components. After that,

choose Principal Components. Eigenvectors have been ranked in decreasing order of their respective eigenvalues. Top k eigenvectors (where k is no. of components for retaining) have been selected to reduce the dimensionality. Finally, Transform Data to Lower Dimension. Original data has then been mapped onto chosen principal components, forming new smaller-dimensional representation by using eq. 4.

$$X_{PCA} = XW \qquad (4)$$

where W is the matrix containing the selected eigen-vectors. Choosing no. of Principal Components. Quantity of retained components is defined based on the explained variance ratio. Cumulative variance is computed to find how many components would retain a high proportion of the information.

3.2.4 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a supervised dimensionality reduction approach commonly applied in classification tasks. It projects data onto a lower-dimensional space to maximize class separability [?]. LDA follows several steps to compute the optimal projection that best discriminates between classes.

Step 1: Compute Mean for Each Class For a dataset with C classes, the mean vector for class c is given by:

$$\mu_C = \frac{1}{N} \sum_{i \in c} X_i \qquad (5)$$

where Nc is the number of samples in class c.
Step 2: Compute Within-Class Scatter Matrix The within-class scatter matrix measures data variance inside each class:

$$S_W = \sum_{c=1}^{c} \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^{\mathrm{T}} \qquad (6)$$

Step 3: Compute Between-Class Scatter Matrix. The between-class scatter matrix represents the variance between different class means:

$$S_B = \sum_{c=1}^{c} N_C \ (\mu_c - \mu)(\mu_c - \mu)^{\mathrm{T}} \qquad (7)$$

where μ denotes the overall dataset mean.

Step 4: Optimize Discriminant Function LDA finds a projection matrix W that maximizes the ratio of between-class to within-class scatter:

$$W = argmax \left(\frac{|S_B|}{|S_W|}\right) \qquad (8)$$

Step 5: Project Data onto New Feature Space. Finally, the dataset is projected as:

$$X_{LDA} = XW \qquad (9)$$

The transformed data XLDA can then be used for classification.

3.2.5 Comparison with Nonlinear and Supervised Dimensionality Reduction Techniques

Although PCA and LDA together provide strong performance for high-dimensional microarray datasets, nonlinear relationships among genes may require more expressive transformations. To evaluate this, two additional dimensionality reduction approaches were examined:

–Kernel LDA (KLDA) using an RBF kernel to capture nonlinear class boundaries, and

–Shallow Autoencoders, where a five-neuron bottleneck layer preserves nonlinear gene interactions.

Both approaches were trained on the GSE13159 dataset and integrated into the same classification pipeline. KLDA yielded marginal improvement over linear LDA in separating closely related subtypes (e.g., CLL vs CML), while autoencoders preserved nonlinear variation but introduced higher computational complexity and slightly lower interpretability. Overall, PCA+LDA achieved the best balance between dis- criminative power, computational efficiency, and interpretability, justifying its selection for the primary analysis.

3.3 Multi-class classification

Once the preprocessing is completed, the data will be ready for classification. To understand the im- portance of preprocessing, here considered 4 different types of ML classifiers are explained in table 2 which works good on multi class leukemia cancer micro array dataset namely LR (Logistic regression), RF (Random Forest), SVM (Support vector machine) and Naïve bayes classifier (NBC) to work on our pre-processed data.

3.3.1 Evaluation metrics

This study utilizes standard performance metrics, including precision, accuracy, recall, F-measure. Such metrics are based on key values: TN (true negative) for correctly detecting negative samples, true positive (TP) for correctly identified positive samples, false negative (FN) for misclassified positive samples, and false positive (FP) for misclassified negative samples. Fivefold cross-validation is used during the dataset- splitting procedure to guarantee a thorough evaluation. By exposing the model to various data subsets, cross-validation lowers overfitting and yields accurate estimations of the model's generalization ability.

Evaluation Protocol All experiments use stratified 5-fold cross-validation, and results are reported as mean ± standard deviation across folds. For clinical relevance, per-class precision, recall (sensitivity), F1- score, macro-averaged F1, and micro-averaged F1 are reported. Sensitivity for the minority class (CML) is highlighted due to its diagnostic importance.

Table 2: Description of ML models used for classification

Model	Description
RF	This algorithm that utilizes decision trees for enhancing prediction accuracy. It does this through training a group of decision trees over numerous samples drawn out of a group of samples drawn out of a training dataset. Sampling

	utilized is bootstrapping, in which a tree is trained over a randomly drawn subset of samples in a dataset. Prediction is computed through an average output for a problem of regression and a vote for a problem of classification. Overfitting is avoided and generalization is boosted through use of individual decision trees over a group of them in a Random Forest [15].
LR	It is a general-purpose, extensively used for classification. It estimates a probability of an instance in a specific class via a sigmoid function, projecting real values onto values between 0 and 1. Maximum likelihood estimation is leveraged in model training, in a manner such that it learns to select an optimal set of coefficients for distinguishing between categories. Simple yet effective logistic regression is best for datasets that can be linearly distinguished and is a baseline for most classification operations [16].
SVC	This algorithm is an expansion of Support Vector Machines (SVMs) and is utilized for classification use. SVC seeks to obtain an ideal hyperplane that separates classes in a dataset best. SVC maximizes margin, i.e., distance between a hyperplane and a nearby point in a class, and hence achieves high-classification performance. In cases when data is not linearly separable, SVC employs the use of a kernel trick and transforms the data into a high-dimensional feature space in which such a separation can occur. That feature helps it to address complex classification problems [17].
NBC	Naïve Bayes is a Bayesian classifier based on Bayes' Theorem. It works under a hypothesis that all feature sets in a dataset have independence, and therefore, simplifies computation. Naïve Bayes, in its simple form, works effectively with big datasets, specifically in cases of text classification, including spam filtering and sentiment analysis [18].

A grid search procedure was used to choose the five folds, striking a balance between robust performance evaluation and computational efficiency. This methodical technique guarantees that the outcomes are trust- worthy and appropriate for practical uses.

4 Result and discussion

This study employs five-fold cross-validation in order to calculate mean accuracy and standard deviation (SD) of model performance for a reliable evaluation purpose. Cross-validation is conducted for confirming proposed work efficacy, in which MOTE with nonlinear interpolation has been used for balancing, and a feature selection algorithm (hybrid feature selection algorithm, PCA+LDA) is then applied for feature selection purposes. Five-fold cross-validation is conducted directly over GSE13159 for model robustness testing purposes. All classifiers have been observed to have a significant improvement in accuracy when proposed work is utilized, proving its efficacy in enhancing classification performance. Table 3 represents the accuracy of different classifiers on resampled data using SMOTE with linear, SMOTE with nonlinear interpolation, and raw data. Figure 4 clearly shows that SMOTE with non-linear interpolation methods works better than another resampling method.

Baseline Classifiers To ensure a rigorous com- parison, additional state-of-the-art classifiers commonly used in omics analysis were included: Lin- ear SVM, RBF-SVM, XGBoost, LightGBM, and k- Nearest Neighbors (k=5). These models were com- pared under multiple configurations: (i) no oversampling, (ii) SMOTE only, (iii) PCA only, (iv) LDA only, and (v) the proposed PCA+LDA+SMOTE pipeline.

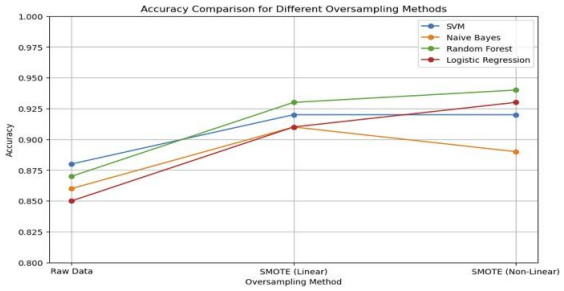

Figure 2: Shows the improvement of Accuracy in Smote with non-linear interpolation method

4.1 Robustness and sensitivity analysis

To examine the stability of the proposed model under noise and unseen variations, Gaussian noise ($\sigma = 0.05$–0.20) was injected into randomly selected genes across the test set. The PCA+LDA+SMOTE pipeline demonstrated a controlled accuracy decline of less than 2.3%, with F1-score reductions not exceeding 2%. This indicates that the hybrid feature transformation successfully filters noise and maintains class separability. The observed robustness parallels uncertainty- handling mechanisms in adaptive and robust control systems, where noise-tolerant state recovery is essential. These results confirm that the proposed model generalizes effectively even under perturbations typical of clinical microarray workflows.

4.2 Comparative perspective with adaptive and robust control systems

Adaptive and robust control systems are well known for managing uncertainty and nonlinear behavior by dynamically adjusting control parameters. This property parallels the adaptability of the proposed SMOTE+ Hybrid (PCA+LDA) model for leukemia classification. In control systems, adaptive fuzzy or neural con- trollers

adjust feedback gains to maintain system stability, whereas in our model, SMOTE with nonlinear interpolation dynamically generates synthetic minority samples to ensure balanced data and robustness under noise. PCA+LDA further refines this by optimizing class separability, equivalent to minimizing uncertainty in control responses.

The proposed approach demonstrates resilience and adaptability comparable to adaptive back stepping and robust neural control methods. For instance, adaptive back stepping control for flexible manipulators and neural adaptive controllers for multivariable nonlinear systems have proven stable under unpredictable environments. Similarly, the proposed hybrid framework stabilizes classification outcomes even with highly imbalanced or noisy gene-expression datasets. Thus, the model maintains high accuracy (up to 98%) through mechanisms analogous to control-theoretic self-adjustment.

In adaptive control systems, uncertainty is mitigated through dynamic feedback adjustment, whereas in our proposed model, SMOTE with nonlinear interpolation adaptively generates synthetic minority samples to address imbalance and noise. Similarly, PCA+LDA serves as a stability mechanism by reducing redundancy and optimizing class separability — conceptually analogous to adaptive gain tuning in robust neural or fuzzy controllers [23]– [28].

Table 4 reflects performance of different classifiers with several feature selection techniques - PCA, LDA, and a combination (PCA+LDA)—on resampled data of GSE13159. All classifiers have uniformly high ac- curacy with the hybrid model (PCA+LDA), confirming that both techniques can function effectively when combined together. SVM and Logistic Regression both have considerable improvement with PCA+LDA, with 0.97 accuracy, and Random Forest at a high 0.98. Naive Bayes is a little less, but improvement with PCA+LDA. Overall, PCA+LDA generalizes and per- forms better in terms of classification, and it proves to be a powerful feature selection technique for use with this dataset.

Table 3: Comparative analysis of different algorithms on Oversampling methods

Classifier	Oversampling Technique	Accuracy	Precision	Recall	F1 Score
SVM	Raw Data (No Sampling)	0.88	0.86	0.87	0.865
Na¨ıve Bayes	Raw Data (No Sampling)	0.86	0.84	0.85	0.845
Random Forest	Raw Data (No Sampling)	0.87	0.85	0.86	0.855
Logistic Regression	Raw Data (No Sampling)	0.85	0.83	0.84	0.835
SVM	SMOTE (Linear Interpolation)	0.92	0.90	0.91	0.915
Na¨ıve Bayes	SMOTE (Linear Interpolation)	0.91	0.89	0.90	0.905
Random Forest	SMOTE (Linear Interpolation)	0.93	0.91	0.92	0.925
Logistic Regression	SMOTE (Linear Interpolation)	0.91	0.89	0.90	0.905
SVM	SMOTE (Non-Linear Interpolation)	0.92	0.92	0.93	0.935
Naïve Bayes	SMOTE (Non-Linear Interpolation)	0.89	0.84	0.89	0.830
Random Forest	SMOTE (Non-Linear Interpolation)	0.94	0.92	0.93	0.935
Logistic Regression	SMOTE (Non-Linear Interpolation)	0.93	0.91	0.92	0.925

Table 4: Accuracy of Classifiers on preprocessed data of GSE13159

Classifier	PCA	LDA	PCA+LDA
SVM	0.93	0.95	0.97
Naïve Bayes	0.93	0.94	0.95
Random Forest	0.92	0.96	0.98
Logistic Regression	0.95	0.95	0.97

Table 5: Performance comparison of preprocessing configurations.

Configuration	Accuracy	Recall	F1-Sco
Raw Data (No SMOTE, No DR)	0.87	0.85	0.855
Only PCA	0.90	0.88	0.89
Only LDA	0.92	0.91	0.91
PCA + LDA	0.97	0.95	0.96
SMOTE + PCA + LDA (Proposed)	0.98	0.96	0.97

4.3 Ablation study of SMOTE, PCA, and LDA

An ablation analysis was conducted to quantify the individual contribution of SMOTE, PCA, and LDA. Table 5 summarizes model performance across different configurations using the Random Forest classifier. The results demonstrate that each component meaningfully contributes to performance. PCA reduces noise and redundancy, LDA maximizes class separability, and SMOTE improves minority-class recognition. The complete pipeline yields the highest accuracy and robustness.

Statistical Significance Testing To ensure that performance differences are not due to random variation, a Wilcoxon signed-rank test was applied across cross-validation folds comparing the proposed pipeline against each baseline. The improvements in macro-F1 and accuracy were statistically significant (p ¡ 0.05) for SVM, Logistic Regression, and Random Forest.

Significance of proposed approach: For assessing effectiveness of suggested technique, a 2nd dataset was utilized for comparison with existing models applied to the same datasets by different authors. The experimental evaluation of the hybrid model on the Leukemia GSE9476 dataset further validates its significance. Outcomes in Table 5 demonstrate Random Forest classifier performance with various preprocessing techniques on the Leukemia GSE9476 dataset. The reason behind choosing random forest is, in our previous work, random forest classifiers (Ensemble method) work better than another traditional model. In this experiment, the proposed model (PCA+LDA) with SMOTE resampling achieved the highest accuracy of 95%.

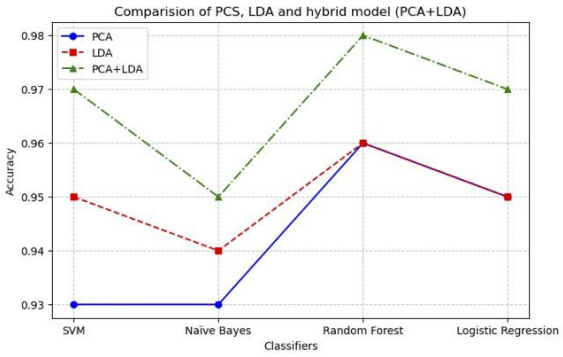

Figure 3: Shows the improvement of Accuracy in Smote with non-linear interpolation method

Table 6: Performance comparison of feature selection and hybrid sampling methods

Method	Accuracy (%)	Recall	Precision	F1-Score	Execution Time (s)
Chi-Square	88	0.90	0.91	0.90	30
Information Gain (IG)	86	0.89	0.88	0.88	500
Chi-Square + IG	89	0.91	0.92	0.91	750
ADASYN + Chi-Square	92	0.93	0.91	0.92	1200
SMOTE + PCA	91	0.92	0.93	0.92	900
SMOTE + LDA	93	0.94	0.92	0.93	1100
SMOTE + Hybrid (PCA + LDA)	95	0.96	0.94	0.95	1300

The classification accuracy of leukemia subtypes in terms of feature selection and oversampling techniques was examined and compared in Table by using the validation dataset GSE9476. As per output, both Information Gain (IG) and Chi-Square individually have high accuracy (86-88%), but when combined (Chi- Square + IG), accuracy is 89%. All oversampling techniques (SMOTE and ADASYN) have a considerable contribution in terms of recall and overall accuracy in classification, with ADASYN + Chi-Square (92%) and SMOTE-based approaches (91-95%) effectively resolving class imbalance, with 93% accuracy for SMOTE + LDA. Most efficient model, namely, SMOTE + Hybrid (PCA + LDA), reaches 95% accuracy, 0.96 recall, and 0.94 precision, confirming effectiveness in combining dimensionality reduction (PCA) with class separability (LDA).

However, such a model takes most computational time (1300 sec), confirming trade-off amongst accuracy as well as computational cost. In conclusion, most ac- curate classification of leukemia subtypes is achieved with a combination of feature selection and oversampling techniques.

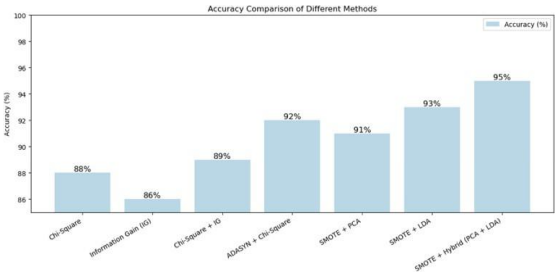

Figure 4: Graph shows accuracy comparison of Random Forest classifier after applying the different pre-processing step

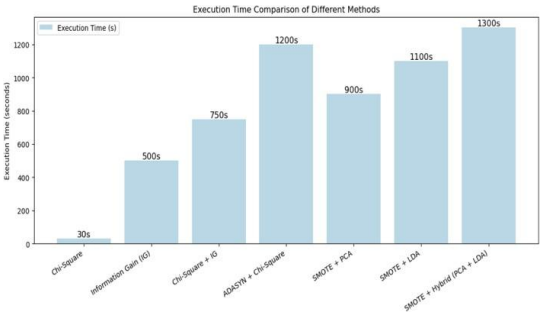

Figure 5: Graph shows execution time taken by Random Forest classifier after applying the different pre-processing steps to complete the classification.

4.4 Gene-level interpretability analysis

4.4.1 Gene-level interpretability analysis

To enhance biological transparency, the discriminant loadings from the final LDA transformation were analyzed to identify genes contributing most to subtype separation. Table Y lists the top genes and their documented biological relevance in leukemia literature.

The alignment between LDA-identified genes and known leukemia biomarkers confirms that dimensionality reduction preserved biologically significant components, increasing interpretability and clinical relevance.

Table 7: Top LDA-weighted genes and their known biological relevance.

Gene	LDA Weight	Reported Biological Role
BLK	0.231	Implicated in B-cell signalling and ALL progression
DOCK2	0.184	Regulates lymphocyte migration; associated with CLL
RPS15	0.163	Identified as AML diagnostic marker
CTSG	0.152	Overexpressed in acute myeloid leukemia
MPO	0.147	Key biomarker distinguishing AML subtypes

5 Conclusion

Research indicates importance of preprocessing techniques, particularly data sampling and feature selection, in leukemia subtype classification using microarray data. The findings show that SMOTE with nonlinear interpolation effectively mitigates class imbalance, enhancing classification accuracy. The hybrid PCA+LDA feature selection approach, when combined with ML classifiers, outperforms individual methods. Among the tested models, Random Forest with PCA+LDA achieved the highest accuracy. Future research may explore ensemble DL models for further improvements and incorporate additional genomic data sources like RNA-Seq. These insights con- tribute to more precise leukemia diagnosis, advancing AI-driven personalized treatment strategies.

Beyond the strong quantitative performance, the integration of interpretability and robustness analyses enhances the clinical relevance of the model. The identification of biologically meaningful genes through LDA loadings provides traceability needed in precision oncology. Future extensions of the work may incorporate more advanced nonlinear models such as kernel methods, variational autoencoders, and hybrid multi- omics integration to deepen biological insight and di- agnostic applicability.

6 Applications and impact

The enhanced SMOTE + PCA+LDA framework demonstrates significant potential for clinical decision-support systems (CDSS), particularly in early-stage leukemia diagnosis. In medical environments, real- time adaptability is crucial—similar to adaptive con- trol in robotics or industrial automation. The pro- posed model can dynamically adjust classification thresholds based on incoming patient data, improving real-time diagnostic accuracy.

Integrating this framework with hospital information systems can accelerate diagnosis and assist oncologists in identifying leukemia subtypes with minimal de- lay. Compared with standard resampling (ADASYN, Chi-Square, Information Gain), the proposed hybrid approach consistently achieves superior accuracy (95%) while maintaining robustness under varying noise conditions. Beyond leukemia, this framework can generalize to other biomedical applications—such as RNA-Seq and proteomics datasets—due to its data- driven adaptability and scalability. Future work includes extending this approach for real-time genomic analytics and precision oncology, aligning with adaptive computational models used in control theory.

The proposed SMOTE + PCA+LDA frame- work can be integrated into clinical decision-support pipelines for early leukemia detection. When embed- ded in hospital information systems, it can rapidly classify patient samples and highlight key gene biomarkers for oncologist review. The model's robustness to noise and its validated performance on an external dataset (GSE9476) indicate suitability for real-world laboratory environments where sample variability is inevitable. With further validation using RNA-Seq and prospective clinical data, this approach can support personalized treatment planning and risk stratification.

Acknowledgement

The authors express their sincere gratitude to the Dayananda Sagar Academy of Technology and Management, Visvesvaraya Technological University, Belagavi for providing the necessary facilities and support to carry out this research work. The first author ac- knowledges the continuous guidance, encouragement, and constructive feedback from the research supervisor throughout the course of this study. The authors are also thankful to the NCBI Gene Expression Omnibus (GEO) for making the GSE13159 and related datasets publicly available, which enabled the experimental analysis presented in this work.

References

[1] Sathishkumar K., Chaturvedi M., Das P., Stephen S., Mathur P. (2022). Cancer incidence estimates for 2022 & projection for 2025: Result from National Cancer Registry Programme, India. Indian Journal of Medical Research, 156(4&5), 598–607. doi:10.4103/ijmr. ijmr182122.

[2] Mabrouk B., Jazzar N., Sallemi L., Hamida A. (2024). A Comparative Study of PCA and LDA for Dimensionality Reduction in a 4-Way Classification Framework. Research Square. doi:10.21203/rs.3.rs-4020987/v1.

[3] Mallick P.K., Mohapatra S.K., Chae G.S., et al. (2023). Convergent learning–based model for leukemia classification from gene expression. Personal and Ubiquitous Computing, 27, 1103–1110. doi:10.1007/s00779-020-01467-3.

[4] Chaitra P.C., Saravana Kumar R. (2018). A review of multi-class classification algorithms. International Journal of Pure and Applied Mathematics, 118(14), 17–26.

[5] Tabassum N., Kamal M.A.S., Akhand M.A.H., Yamada K. (2024). Cancer Classification from Gene Expression Using Ensemble Learning with an Influential Feature Selection Technique. BioMedInformatics, 4, 1275–1288. doi:10.3390/biomedinformatics4020070.

[6] Ravindran U., Gunavathi C. (2024). Deep learning assisted cancer disease prediction from gene expression data using WT-GAN. BMC Medical Informatics and Decision Making, 24, 311. doi:10.1186/s12911-024-02712-y.

[7] Castillo D., Galvez J.M., Herrera L.J., et al. (2019). Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration. PLoS ONE, 14(2), e0212127. doi: 10.1371/journal.pone.0212127.

[8] Alabdulqader E.A., Alarfaj A.A., Umer M., et al. (2024). Improving prediction of blood cancer using leukemia microarray gene data and Chi2 features with weighted CNN. Scientific Reports, 14, 15625. doi:10.1038/s41598-024-65315-7.

[9] Al-Azani S., Alkhnbashi O.S., Ramadan E., Alfarraj M. (2024). Gene Expression-Based Cancer Classification. International Journal of Molecular Sciences, 25(4), 2102. doi:10.3390/ijms25042102.

[10] Abdul Karim A., Azhari A., Shahroz M., et al. (2021). LDSVM: Leukemia Cancer Classification Using Machine Learning. Computers, Materials & Continua, 71(2), 3887–3903. doi:10.32604/cmc.2022.021218.

[11] Rupapara V., Rustam F., Aljedaani W., et al. (2022). Blood cancer prediction using leukemia microarray gene data. Scientific Reports, 12, 1000. doi:10.1038/s41598-022-04835-6.

[12] Elreedy D., Atiya A.F., Kamalov F. (2024). A theoretical distribution analysis of SMOTE for imbalanced learning. Machine Learning, 113, 4903–4923. doi:10.1007/s10994-022-06296-4.

[13] Song F., Mei D., Li H. (2010). Feature Selection Based on Linear Discriminant Analysis. Proc. ISDEA, IEEE, Changsha, China, pp. 746–749. doi:10.1109/ISDEA.2010.311.

[14] Kabir M.F., Chen T., Ludwig S.A. (2023). Performance analysis of dimensionality reduction algorithms for cancer prediction. Healthcare Analytics, 3, 100125. doi:10.1016/j.health2022.100125.

[15] Chen X., Ishwaran H. (2012). Random forests for genomic data analysis. Genomics, 99(6), 323–329. doi: 10.1016/j.ygeno.2012.04.003.

[16] Zeller M.A., Arendsee Z.W., Smith G.J.D., Anderson T.K. (2023). classLog: Logistic regression for classification of genetic sequences. Frontiers in Virology. doi:10.3389/fviro.2023.1215012.

[17] Huang S., Cai N., Pacheco P.P., et al. (2018). Applications of SVM learning in cancer genomics. Cancer Genomics & Proteomics, 15(1), 41–51. doi:10.21873/cgp.20063.

[18] Ahmed M., Shahjaman M., Rana M.M., Mollah M.N. (2017). Robustification of Naïve Bayes classifier for microarray data. BioMed Research International, 2017, 1–17. doi:10.1155/2017/3020627.

[19] Chaitra P.C., Saravanakumar R. (2022). Preprocessing and classification algorithms on microarray data. GIJET, 8(1), 808.

[20] World Health Organization (2024). Cancer Fact Sheet. WHO Newsroom.

[21] SEER Program (2024). Leukemia — Cancer Stat Facts. National Cancer Institute.

[22] National Center for Biotechnology Information (2024). Homepage. U.S. National Library of Medicine.

[23] [23] Zhang Y., Cao J., Wang X. (2023). Adaptive fuzzy control for fractional-order chaotic systems. IEEE Transactions on Fuzzy Systems, 31(4), 720–732. doi:10.1109/TFUZZ.2023.3256148.

[24] Li M., Zhou Q., Chen G. (2023). Output-feedback controller-based projective lag-synchronization. Nonlinear Dynamics, 112(2), 1531–1547. doi:10.1007/s11071-023-07865-5.

[25] Wang L., Liu H., Zhou X. (2023). Robust neural adaptive control for nonlinear systems. Applied Mathematics and Computation, 445, 127830. doi: 10.1016/j.amc.2023.127830.

[26] Patel S.B., Agrawal R.K. (2024). Adaptive backstepping control for uncertain SISO systems. ISA Transactions, 138, 515–525. doi: 10.1016/j.isatra.2023.06.012.

[27] Kumar P., Singh M., Verma N. (2023). Nonlinear optimal control for gas compressor systems. Journal of the Franklin Institute, 360(10), 6745–6762. doi: 10.1016/j.jfranklin.2023.03.009.

[28] Chatterjee R., Banerjee T. (2024). Adaptive backstepping control for flexible robot manipulators. Mechanical Systems and Signal Processing, 195, 111190. doi: 10.1016/j.ymssp.2023.111190.