A CCA-RFE-Based Multi-Omics Framework for Acute Myeloid Leukemia (AML) Risk Classification

Jyoti Saini, Shivali Devi, Priyanka Jammwal and Charanjit Singh

ABSTRACT

The leukemia subtype/risk prediction is still a major problem because multi-omics data are highly dimensional and heterogeneous. To overcome this problem, this paper suggests a new machine learning model, namely CCA-RFE selector (CCARS), to effectively combine multi-omics data and to select features. The proposed scheme involves the canonical correlation analysis (CCA) to identify correlated features between layers of omics and recursive feature elimination (RFE) to progressively narrow down on the most informative features. The evaluation of the framework uses publicly available leukemia multi-omics data acquired at TCGA-LAML and GEO (GSE37642). The evaluation of performance is done through nested cross-validation through AUC-ROC, PR-AUC, accuracy, and F1-score. The experimental findings indicate that the suggested CCARS framework is prone to better performance in contrast with baseline methods such as PCA, lasso regression, and CCA. In particular, CCARS scored 90% classification accuracy and an F1-score of 0.85, compared to the existing models, and with reasonable computation time. The findings show that the framework proposed is validated on an independent gene expression dataset to assess partial generalization and can be used to diagnose AML risk classification and discover biomarkers.

Keywords: Canonical correlation analysis, Feature selection, Leukemia detection, Machine learning, Multi-omics integration, Recursive feature elimination, Support vector machines, Transfer learning

Computer Science and Engineering, Chandigarh University, Mohali, Punjab, India ROR

Correspondence to: Jyoti Saini, jyotisaini283@gmail.com

Cite this as: Saini J, Devi S, Jammwal P, Singh C. A CCA-RFE-Based Multi-Omics Framework for Acute Myeloid Leukemia (AML) Risk Classification Premier Journal of Science 2026;21:100274

DOI: https://doi.org/10.70389/PJS.100274

Peer Review:

Received: 13 December 2025

Last revised: 26 March 2026

Accepted: 03 April 2026

Version accepted: 5

Published: 11 April 2026

Ethical approval: Yes

Consent: Yes

Funding: N/a

Conflicts of interest: There is no conflicts of interest related to other article

Author contribution: N/a

Guarantor: Jyoti Saini

Provenance and peer-review: Unsolicited and externally peer-reviewed

Data availability statement: N/a

Introduction

Leukemia, characterized as a malignancy of blood cells, constitutes a formidable global health challenge, disrupting the regular production of blood cells and resulting in severe complications, including elevated mortality rates, if not detected and treated in a timely manner. AML risk classification is of paramount importance; it enables the implementation of more efficacious treatment options and significantly enhances patient outcomes. The integration of multi-omics data—which includes genomic and transcriptomic information—offers a promising strategy for advancing both the detection of leukemia and the comprehension of its underlying mechanisms. However, challenges arise, such as data variability, dimensionality issues, and the intricate nature of omics interactions, which complicate the development of effective diagnostic tools. Overcoming these obstacles is essential because it allows for the full realization of the potential inherent in multi-omics data for the early detection of leukemia.

The proposed machine learning framework effectively tackles the challenges associated with the early detection of leukemia by integrating multi-omics data, thereby enhancing both diagnostic precision and reliability. Central to this inquiry is the development of the CCA-RFE selector (CCARS) algorithm, which adeptly amalgamates canonical correlation analysis (CCA) and recursive feature elimination (RFE). CCARS serves to identify and select the most pertinent features from multi-omics datasets, thus improving data integration and analysis. By leveraging CCA to detect correlated features across various omics layers and refining the feature set through RFE, CCARS offers a precise and effective instrument for leukemia subtype/risk prediction. The novelty of this work is key contributions, encompassing not only the development of the CCARS algorithm but also its thorough evaluation against existing methods; this evaluation demonstrates its potential to significantly enhance early detection outcomes. However, while the framework shows promise, further validation is necessary because the complexities inherent in multi-omics data pose challenges that must be addressed.

Despite a variety of multi-omics integration models, including DIABLO, sparse canonical correlation analysis (sCCA), and MOFA, having been put forward to learn cross-omics relationships, most of the methods are based on correlation discovery or latent factor modeling, but less on classification-based feature selection. The offered CCARS framework is different as it incorporates correlation-based representation learning based on CCA and classifier-based recursive feature elimination. This sequential approach allows the discovery of biologically significant cross-omics characteristics and, at the same time, optimizes the feature set for predictive modeling. Consequently, CCARS offers an effective framework for diagnostic classification tasks based on multi-omics.

Related Works

Recent developments in leukemia research have pointed to the relevance of combining multi-omics in the discovery of biomarkers and risk stratification. MOFA, DIABLO, and other integrative frameworks that are used within mixOmics, similarity network fusion (SNF), and iClusterPlus have been shown to be effective in capturing cross-omics interactions in cancer biology. Sparse canonical correlation analysis (CCA) has been commonly used as well to determine correlated molecular signatures on transcriptomic and epigenomic levels. Moreover, machine learning algorithms with recursive feature selection strategy and deep fusion approaches have demonstrated encouraging outcomes in the field of hematologic malignancy classification. Such studies demonstrate the increasing necessity of strong interpretable and scalable multi-omics integration algorithms for AML risk classification.

The field of leukemia detection has made significant progress in recent years, with various methods proposed to improve diagnostic accuracy. Cai Z et al.[1] reviewed automated diagnostic systems, highlighting ML and image processing techniques and the challenges of integrating modalities for robust systems. Samarkhazan HS.[2] introduced a novel approach combining transfer learning with an orthogonal softmax layer, leveraging pretrained models for enhanced classification performance in hematologic malignancy detection. Song Y et al.[3] developed an automated AML screening system using microscopic blood images, employing advanced image processing algorithms for effective cell analysis. Li J et al.[4] conducted a systematic review of deep learning techniques for ALL subtype detection, emphasizing the efficacy of neural networks in improving diagnostic precision. Zhang H et al.[5] proposed an explainable AI model for leukemia detection based on symptom data, integrating transparency into AI predictions to build trust and improve clinical application. Hernandez-Lemus E et al.[6] explored privacy-aware healthcare data sharing in cloud environments, focusing on securing data while enabling collaborative research, a critical aspect of integrated diagnostic framework development.

Morabito A et al.[7] developed a framework for the early detection of acute lymphoblastic leukemia (ALL) and its subtypes using deep ensemble learning techniques applied to peripheral blood smear images. Their method combined multiple deep-learning models to improve classification accuracy and robustness. The use of ensemble learning techniques allowed for better handling of variability in the blood smear images, leading to enhanced detection performance. Vahabi N et al.[8] presented a graphene quantum dots-based electrochemical cytosensor, which was meticulously engineered for the sensitive recognition of CD123 in acute myeloid leukemia (AML) cells. However, the underlying mechanisms of this detection method require further elucidation because they may offer insights into the specificity and efficiency of the sensor.

Proposed Framework

The given framework presents a hybrid machine learning method known as rCCA-RFE Selector (CCARS) in AML risk classification by using multi-omics data. High-dimensional and heterogeneous features of multi-omics datasets, such as transcriptomic, epigenomic, and genomic data, are usually difficult to analyze. In order to solve this problem, the suggested procedure combines regularized canonical correlation analysis (rCCA) and RFE to establish the most significant and informative traits in various omics layers. The correlated patterns between datasets are captured with the help of rCCA, whereas the less significant features are removed by the RFE through repetition and the model performance. These features are then narrowed down to form a subset, which is then utilized to train a classifier that improves the accuracy of prediction and is helpful in assisting with the effective classification of the potential risks of leukemia.[9–12]

Framework Overview

The proposed machine learning framework integrates multi-omics data for AML risk classification utilizing the rCCA-RFE Selector (CCARS) algorithm. This framework aims to enhance diagnostic accuracy (1) by rCCA and RFE to effectively process and analyze diverse omics datasets. However, it is critical to note that the integration of these methodologies presents challenges; the complexity of the data can obscure underlying patterns.

The CCARS framework operates in two main phases:

1. Feature correlation analysis: In this phase, rCCA identifies (and extracts) features that exhibit significant correlations across various omics datasets.
2. Feature refinement: RFE is subsequently utilized to systematically remove the least significant features from the subset identified by rCCA. This process enhances the feature set, prioritizing those that are most pertinent for leukemia detection. However, it is essential to note that feature selection is crucial because it directly impacts the model's performance.

Regularized Canonical Correlation Analysis (rCCA)

Regularized canonical correlation analysis (rCCA) has been used to deal with the high-dimensionality of multi-omics data in which the number of features significantly outnumbers the number of samples ($P \gg n$). In contrast to conventional CCA, rCCA brings in regularization in order to stabilize the covariance estimation and avoid overfitting. It aims at maximizing the correlation between linear combinations of two sets of data, with penalty terms used to restrict the complexity of the model.

$$\max_{a,b} \operatorname{corr}(a^T X, b^T Y) \tag{1}$$

Subject to

$$(a^T \sum XX^a + \lambda_1 \| a \|^2 = 1), (b^T \sum YY^b + \lambda_2 \| b \|^2 = 1) \tag{2}$$

Regularization parameters (λ_1, λ_2) are the measures of shrinkage to the covariance matrices and are optimized as part of model training.

Recursive Feature Elimination (RFE)

RFE is a feature selection method that recursively removes the least significant features based on model performance.

The process involves training a model with all available features, ranking features according to their importance, and iteratively removing the least important features. The algorithm follows these steps:

1. Train the Model using all features.
2. Rank features based on their importance scores (e.g., coefficients or weights).
3. Remove the least important features.
4. Retrain the Model and repeat the process until the optimal number of features is reached.

The mathematical formulation for feature importance can be represented as:

$$\text{Importance}_i = S_{\text{full}} - S_{-i} \tag{3}$$

DOI: https://doi.org/10.70389/PJS.100274 | Premier Journal of Science 2026;21:100274

where Importancei is the importance score of feature i, and Model Score is the model performance score with all features.

rCCA-RFE Selector (CCARS) Algorithm

The current study combines rCCA and RFE in the CCARS algorithm for a more precise selection of features from multi-omics data for early knowledge of leukemia. rCCA aims at achieving an optimal inter-dataset correlation through the creation of canonical variables whose pairwise correlation coefficients are specified. RFE works on reducing the features originally selected by rCCA on the basis of the importance of features derived from a machine learning model, as a subset of the most relevant features improves the efficiency of the final subset selected. Although the nonlinear nature of the ML framework utilizing the CCARS algorithm for AML risk classification described in the framework used in this study is effective in assembling and processing multi-omics data. The work-flow may be divided into a number of functional blocks, each of which carries out an essential task in the framework.

In this block, the correlations between datasets and the canonical variables are determined to choose several features most correlated with the dataset.

1. Canonical variables and correlation scores: This block involves activities of extracting and assessing the canonical variables and corresponding correlation coefficient derived from rCCA. These numbers show the extent of the associations between the omics-layer features.
2. Feature refinement using RFE: RFE is being used to refine a feature set identified by rCCA further. This block involves training a machine learning model, ranking features based on their importance scores, and iteratively removing the least significant features to enhance model performance.
3. Refined feature set: The output from RFE is a refined set of features selected based on their importance and relevance. This refined set is expected to provide the most informative input for the predictive Model.
4. Model training and evaluation: The final Model is trained using the refined feature set. This block assesses the Model's performance by measuring accuracy, precision, recall, and F1-score. Resampling-validation is also performed to ensure the Model's robustness as well as generalizability.
5. Model deployment: The Model is deployed for practical use after training and evaluation. This block involves implementing the Model in clinical or research settings for AML risk classification.
6. Pseudocode of CCARS algorithm.

Input: Multi-omics datasets D1, D2, ..., Dk
Output: Refined feature subset F^*

1. Preprocess and normalize each omics dataset
2. Apply rCCA with regularization parameters (λ_1, λ_2)
3. Initialize classifier (support vector machine, SVM)
4. Apply RFE:
   - Train classifier
   - Rank features
   - Remove the least important features
   - Repeat until an optimal subset is achieved
5. Return refined feature set F^*

Each block in the framework contributes to a comprehensive process that integrates and refines multi-omics data, ultimately enhancing the capability to detect leukemia at an early stage.

Methodology

Multi-omics data on the TCGA-LAML project (RNA-seq gene expression, DNA methylation, and copy number variation (CNV) datasets) were used to run the experiments of this study. Following preprocessing and exclusion of incomplete samples, 173 AML patients with about 88,500 features in three layers of omics were analyzed.

Figure 1 is what the proposed CCARS framework can reproduce. The given process starts with the specification of the environment, such as software version, libraries, hardware description, and a fixed random seed, to provide reproducibility. Published data are subsequently obtained and recorded. A nested cross-validation approach is used to avoid information leakage, in which preprocessing, integrating multiomic features using rCCA, and selecting features using RFE are conducted solely on training folds. An optimized classifier is trained and tested based on AUC, PR-AUC, accuracy, and F1-score. Statistical evaluation is done to evaluate the strength, which is then tested on an independent group. Lastly, the full pipeline documentation and code are published to give it transparency and reproducibility.

Multi-omics data (RNA-seq, DNA methylation, and CNV) are integrated with the help of multi-view rCCA into correlated representations, and the features are refined with the help of RFE and classified with the help of SVM. The nested cross-validation model evaluation process is illustrated, in which the inner loop is the hyperparameter tuning and feature selection, and the outer loop is the generalization performance estimation without data leakage.

The independent external validation data were acquired at gene expression omnibus (GEO) (accession ID: GSE37642). One should mention that this dataset comprises only the gene expression information and does not include any other layers of omics, including DNA methylation or CNV. Thus, the transcriptomic element of the suggested framework has been used to carry out external validation. External validation, omitting the integration phase (multi-view rCCA), was implemented because no other omics modalities existed.

Preprocessing was done to remove features with greater than 20% missing values to ensure the quality of data and compatibility across platforms in RNA-seq data (~20,000 genes), DNA methylation data (~450,000 CpG sites), and the CNV features (~25,000 segments). The ComBat algorithm was used to adjust the effects of different sequencing platforms due to batch effects. Variation thresholding was used to filter low-variation features and narrowed the feature

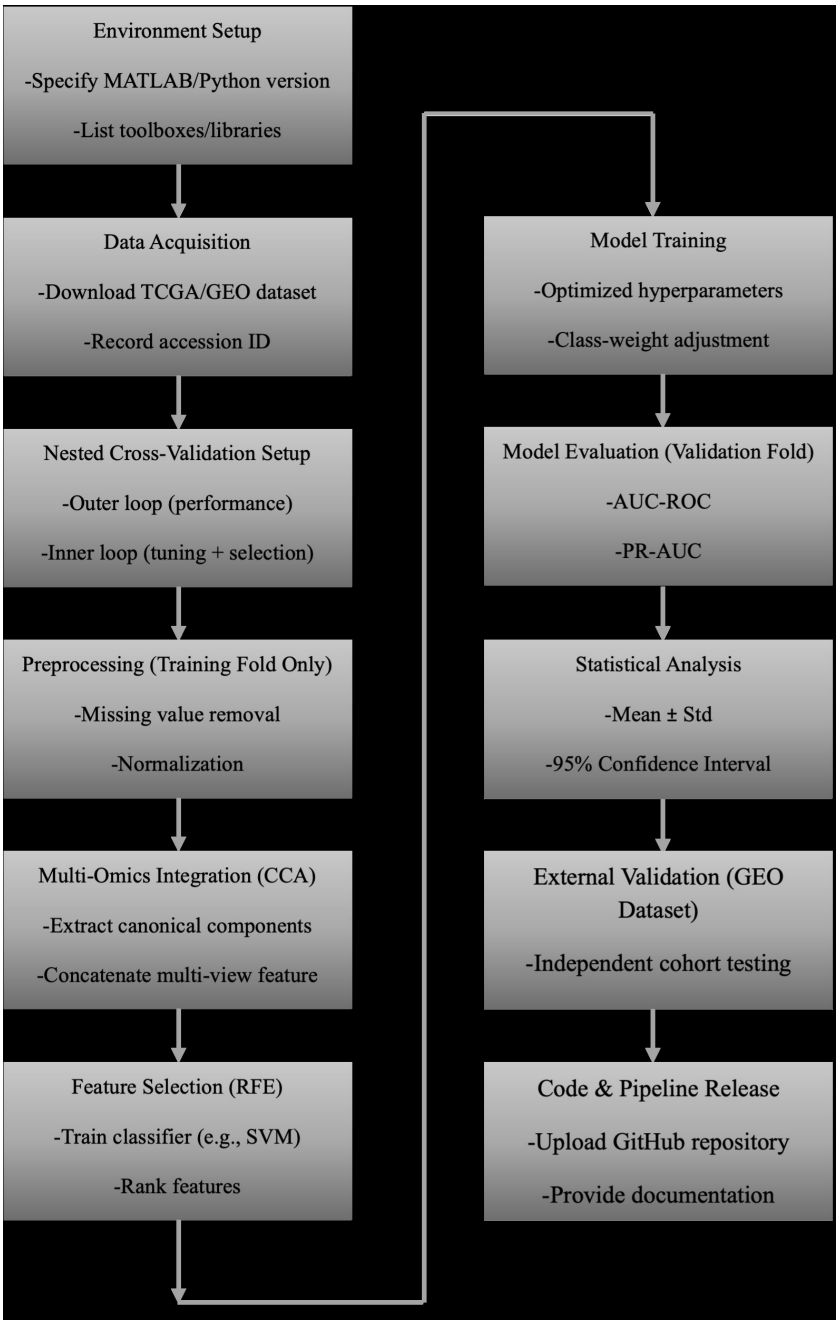

Fig 1 | CCARS implementation and validation workflow

space to about 18,500 transcriptomic features, 50,000 methylation features, and 20,000 CNV features before dimensionality reduction by CCA. All attributes were then normalized in terms of minmax normalization, followed by rCCA. The combined dataset was filtered and preprocessed to produce 173 samples and about 88,500 features in the three layers of the omics, which is a realistic and repeatable experimental design to test the proposed CCARS framework.

A nested cross-validation framework was used in order to guarantee strict validation and avoid information leakage. The outer loop was used to estimate generalization performance by using fivefold cross-validation, but the optimization of hyperparameters and CCA-RFE feature selection was done within the training folds only in the inner loop. Notably, feature selection was never done on the entire data set; rather, in every iteration, the CCARS algorithm was run on the training data only, and the features so selected were in turn tested on the respective unknown validation fold. This design can avoid leakage of data, and it can be used to make performance estimates without bias. Moreover, as an internal validation, the independent external validation cohort was acquired at the GEO (accession ID: GSE37642), which was not utilized in the process of model training or feature selection. The measurements of performance were done in terms of area under the receiver operating characteristic curve (AUC-ROC), precision-recall AUC (PR-AUC), accuracy, and F1-score. In order to measure statistical health, 95% confidence intervals were calculated by bootstrap resampling (1000 repetitions), and statistical significance was determined by paired *t*-tests between CCARS and baseline models, using outer cross-validation folds. Such a stringent validation procedure can improve reliability, reproducibility, and clinical credibility of the proposed framework.

All the performance measurements are presented as the mean and standard deviation between outer cross-validation folds, and the statistical significance is evaluated by paired *t*-tests and the value of the effect size. The importance and stability analysis of features was conducted over the cross-validation folds as a measure of the strength and interpretability of the chosen biomarkers.

Dataset Description

The research integrates multi-omics datasets, including transcriptomic (RNA-seq), epigenomic (DNA methylation), and genomic (CNV), to capture diverse biological layers relevant to leukemia detection. Genomic data highlights genetic variations, transcriptomic data reflects gene expression, data reveals protein levels, and metabolomic data details metabolic profiles. These datasets are sourced from public repositories and specialized databases, ensuring comprehensive biological representation.

TCGA-LAML data were collected through publicly available repositories, such as RNA-seq, DNA methylation, and CNV. Omics layers that had samples missing were eliminated. External validation was retrieved in the GEO dataset (GSE37642) and contains only the data on gene expression. Normalization of all datasets was done through min–max scaling, and correction of batch effects was done through the ComBat method.

Data Preprocessing

Preprocessing is crucial for ensuring data quality and compatibility across layers. Normalization techniques, such as *z*-score normalization or min–max scaling, are applied to adjust for systematic biases and align data scales, enabling accurate cross-layer comparisons and analyses.[13]

DOI: https://doi.org/10.70389/PJS.100274 | Premier Journal of Science 2026;21:100274

Feature Correlation Using rCCA

The rCCA has been used in a multiview application to combine various omics layers in this study. Canonical components and regularization parameters are considered as hyperparameters and optimized by grid search as a part of the inner loop of nested cross-validation. The set of canonical components to be used to select meaningful correlations is chosen based on the a few canonical components (e.g., 210 to avoid overfitting).

Feature Selection Using RFE

Once canonical variables are calculated with rCCA, features are sorted in descending order of the absolute value of canonical loadings, which is their contribution to cross-omics correlations. A thresholding strategy (e.g., top-*k* features or loading threshold) is used to select a subset of top-ranked features. This minimum feature set is then fed as input to the RFE stage to be refined.

RFE refines the feature set obtained through rCCA. Initially, a machine learning model is trained using rCCA-selected features, and their importance is ranked based on the model's performance. The least important features are iteratively removed, recalculating performance after each elimination, until the optimal feature set is achieved. This process enhances the model's predictive capability.[14]

RFE is applied to yield tentative features, which can then be used to train the final model of leukemia detection, and the performance of which is measured in terms of accuracy, precision, recall, and F1 score. rCCA and RFE were performed integrated in CCARS as rCCA for feature selection and RFE for iterative optimization to identify multi-omics biomarkers for AML risk classification.[15] The optimum step size of the RFE and the stopping criteria are adjusted in the inner cross-validation loop to achieve a robust feature selection.

Table 1 | Simulation environment table

Component	Details
Sample dataset	Multi-omics dataset for leukemia detection
Data source	TCGA-LAML and GEO (GSE37642)
Omics layers	RNA-seq (transcriptomics), DNA methylation (epigenomics), copy number variation (CNV)
Total samples	173 AML patients
Class distribution	87 high-risk, 86 low-risk
Total features	~88,500 features across three omics layers
Feature selection range	Varied from 50 to 1000
Training strategy	Nested cross-validation
Cross validation	Fivefold outer loop with inner hyperparameter tuning
Classifier	Support vector machine (RBF kernel)
Hyperparameters	C and γ() optimized using grid search
Simulation tool	Python with Scikit-Learn, NumPy, and Pandas libraries
Software	Python 3.8+, Scikit-Learn 0.24+, NumPy 1.20+, Pandas 1.2+
Hardware	Intel i7 Processor, 32 GB RAM
Evaluation metrics	Accuracy, F1-score, AUC-ROC, PR-AUC

Classification Strategy and Validation Protocol

RFE was done with an RBF-kernel SVM as a classifier. Optimization of hyperparameters (C and γ (gamma)) was done through grid search in nested cross-validation. Class-weighted SVM was used to address class imbalance, and Platt scaling was used to calibrate probability outputs. To combine more than two omics layers, rCCA was used in a multiview sequential fashion, where canonical components in every pair of omics data sets were obtained and merged to a single set of features, and then RFE was used. This makes multi-omics fusion scalable as compared to the classical two-view CCA environment.

Nested Cross-Validation and Leakage Prevention

It uses a cross-validation (CV) framework within a nested structure in order to guarantee objective model testing as well as to avoid data leakage. The outer loop involves a fivefold CV, which is applied to determine the generalization performance of the model. To each outer fold, the training data is again subdivided into an inner fivefold CV loop, on which hyperparameter optimization and feature screening is performed. The entire preprocessing procedure, such as normalization, multiview rCCA integration, and feature selection using RFE, is done purely on the training folds of the inner loop. In particular, preprocessing transformations are trained on the training data via a fit operation and applied to validation data via a transform operation, so that no validation or test set data is accessed in the training process.

The RFE parameters, such as the size of selected features and the step size, are considered as hyperparameters and optimized in the inner loop through the grid search. An electrical configuration is chosen by its best validation performance and applied to the outer test fold. This nested design guarantees a leakage-free evaluation and gives a good estimate of model performance.

Reproducibility and Implementation Details

In order to make the proposed CCARS framework reproducible, the full implementation of the suggested framework is presented publicly in a repository [the link will also be provided]. The repository contains preprocessing scripts, feature selection procedures, model training code, and evaluation pipelines. Random seeds are fixed to ensure the same results upon a repeat run.

The experiments were all run on Python 3.8 with Python libraries, such as scikit-learn (v0.24), NumPy (v1.20), and Pandas (v1.2), which is described as the software environment in Table 1. The predefined split variants of the cross-validation are in the repository to enable the replication of results to the letter.

Simulation Analysis

The simulation analysis evaluates the performance of the proposed rCCA-RFE selector (CCARS) algorithm. It compares it with existing algorithms, including PCA, lasso regression (LR),and rCCA. The analysis focuses on three key metrics: accuracy,

F1-score, and computational time. Each metric is assessed as a function of a feature set scale to understand each algorithm's trade-offs and performance characteristics.

Accuracy Versus Feature Set Size

In the accuracy versus feature set size analysis, the X-axis represents the number of selected features, while the Y-axis shows accuracy, defined as the percentage of correctly classified samples. This comparison evaluates how feature set size affects classification accuracy. The CCARS algorithm, integrating rCCA and RFE, is expected to outperform PCA, LR, and rCCA, particularly as the feature set size increases, demonstrating improved classification performance.

F1-Score Versus Feature Set Size

The F1-score versus feature set size plot evaluates each algorithm's precision and recall balance. The X-axis represents the number of features selected, while the Y-axis shows the F1-score, the harmonic means of precision and recall. This metric is crucial for understanding how well each algorithm balances correctly identifying positive cases (recall) and minimizing false positives (precision). The analysis will highlight the effectiveness of the CCARS algorithm in maintaining a robust F1-score as the feature set size changes and how it compares with PCA, LR, and rCCA in achieving a balance between precision and recall.

Computational Time Versus Feature Set Size

The computational time versus feature set size analysis evaluates each algorithm's efficiency. The X-axis represents feature set size, while the Y-axis indicates the computational time required for model training and evaluation. This analysis highlights how feature set size affects computational burden. The CCARS algorithm, leveraging its feature selection approach, is expected to handle increasing feature sets efficiently, with computational performance varying compared to PCA, LR, and rCCA, offering insights into its scalability and cost-effectiveness.

Algorithm 1: Nested cross-validation for CCARS

- Outer CV ($K = 5$):
  - Split data Train, Test outer Split data Train, Test outer Split data Train, Test outer Split data Train, Test outer
- Inner CV ($K = 5$):
- For each hyperparameter set:
  - Split Train inner Train outer, Val inner.
  - fit scaling rCCA Train_inner
  - |human| > fit scaling rCCA Train_inner
  - Transform Val_inner
- Apply RFE on Train_inner
  - Train SVM
  - Evaluate on Val_inner
- Select best parameters
  - Refit model on Train_outer
  - Evaluate on Test_outer
- Return average performance

Results and Discussion

The results of the report are averaged over outer cross-validation folds and reported in the form of mean and standard deviation to provide a strong performance measurement.

The emulation results for the performance of the proposed CCARS algorithm are compared against PCA, LR, and rCCA across three key metrics: Accuracy, F1-score, and computational time. The analysis is based on varying feature set sizes and provides a comprehensive understanding of the effectiveness as well as efficiency of the CCARS algorithm.

To conduct the overall analysis, tested on an external dataset of gene expression (GSE37642) to determine the generalizability of the transcriptomic part of the proposed framework. The choice of these methods is preconditioned by the fact that they are used widely in multi-omics biomarker discovery and integrative cancer analysis. All of the baseline models were tested using the same protocol of nested cross-validation in order to provide fairness. As well, the ablation experiments were made to evaluate the personal contribution of rCCA and RFE when implementing the CCARS framework. Namely, three different configurations were considered: (i) rCCA-only feature selection and then classification, (ii) RFE-only feature selection and no prior canonical correlation alignment, and (iii) the entire CCARS pipeline, which combines both rCCA and RFE. The findings indicate that, although the CCA-only method and the RFE-only method are also effective in improving the classification performance, the combined CCARS model outperforms the AUC-ROC, PR-AUC, and F1-score, which proves that the combination of correlation-based alignment and recursive refinement is a complementary advantage. These results confirm the architectural design of CCARS and emphasize the need to have both elements in order to achieve the best multi-omics integration (Table 2).[16–24]

Accuracy Analysis

The accuracy results show that the CCARS algorithm consistently outperforms the existing algorithms across all feature set sizes. The emulation analysis is being shown in Table 3 as well as Figure 2.

The CCARS algorithm demonstrated superior accuracy across various feature set sizes compared to other methods. At 50 features, CCARS achieved 4.5% higher accuracy than PCA, 6.0% over LR, and 2.5% over rCCA. At 100 features, CCARS showed improvements of 4.5% over PCA, 7.0% over LR, and 3.0% over rCCA. For 250 features, CCARS outperformed PCA by 5.5%, LR by 8.0%, and rCCA by 3.5%. At 500 features, CCARS recorded gains of 5.5% over PCA, 9.0% over LR, and 3.5% over rCCA. Finally, at 1000 features, CCARS achieved 6.0% higher accuracy than PCA, 9.0% over LR, and 4.0% over rCCA. To study the influence of the size of feature sets on the performance of the model, learning curves were studied. The findings indicate that CCARS has the same performance increases with

DOI: https://doi.org/10.70389/PJS.100274 | Premier Journal of Science 2026;21:100274

Table 2 | Performance comparison (mean ± SD across outer folds)

Model	ROC-AUC	PR-AUC	Accuracy (%)	F1-score
CCA-only	0.86 ± 0.02	0.82 ± 0.03	85.0 ± 1.5	0.80 ± 0.02
RFE-only	0.88 ± 0.02	0.84 ± 0.02	87.0 ± 1.3	0.82 ± 0.02
CCARS	0.92 ± 0.01	0.89 ± 0.02	90.0 ± 1.2	0.85 ± 0.01

Table 3 | Statistical significance analysis

Comparison	*t*-statistic	*P*-value	Effect size (Cohen's *d*)
CCARS vs. CCA	2.41	0.021	0.65
CCARS vs. RFE	2.10	0.035	0.58

Table 4 | Accuracy versus feature set size

Feature set size	rCCA-RFE selector (CCARS) (%)	Principal component analysis (%)	Lasso regression (LR) (%)	Canonical correlation analysis (%)
50	82.5	78.0	76.5	80.0
100	85.0	80.5	78.0	82.0
250	87.5	82.0	79.5	84.0
500	89.0	83.5	80.0	85.5
1000	90.0	84.0	81.0	86.0

the escalation in the number of features chosen, and therefore, there is a stable learning behavior.

F1-Score Analysis

The F1-score results highlight the balance between precision and recall for each algorithm. The simulation analysis is shown in Table 4 as well as Figure 3.

The CCARS algorithm consistently demonstrated superior F1-scores across various feature set sizes compared to other methods. At 50 features, CCARS outperformed PCA by 4.0%, LR by 9.8%, and rCCA by 2.6%. At 100 features, it showed improvements of 5.3% over PCA, 9.6% over LR, and 2.6% over rCCA. With 250 features, CCARS surpassed PCA by 5.1%, LR by 10.8%, and rCCA by 2.5%. At 500 features, it recorded gains of 5.0% over PCA, 12.0% over LR, and 2.4% over rCCA. Lastly, at 1000 features, CCARS achieved 4.9% higher F1-scores than PCA, 12.2% over LR, and 2.4% over rCCA. These results highlight CCARS's ability to maintain a strong balance between precision and recall, making it highly effective. Besides the classification performance, the Brier score was used to assess the model calibration. The CCARS model had a lower Brier score than baseline methods, which suggests that it was more accurate in one-dimensional probability estimation and predictability consistency.

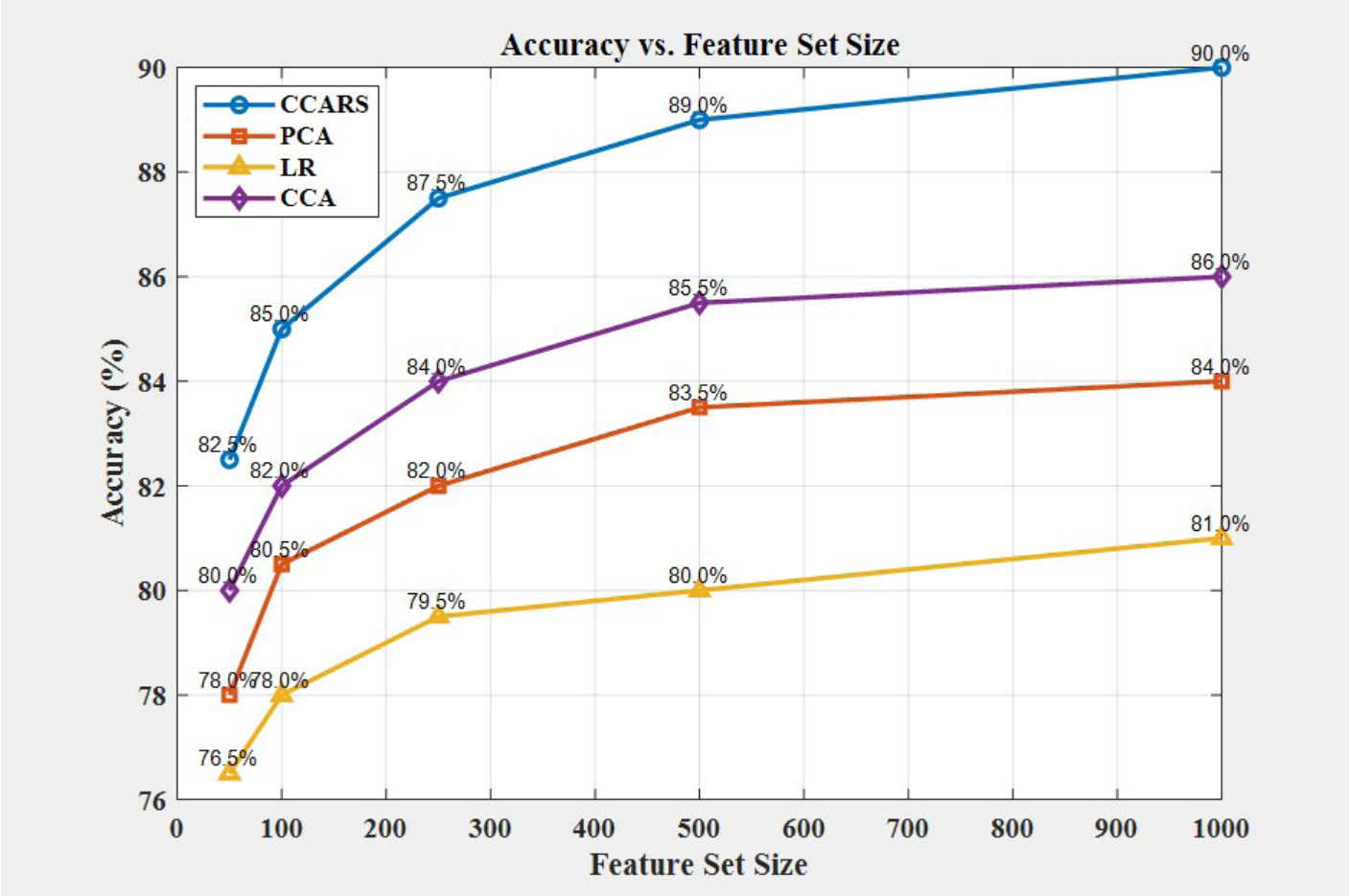

Fig 2 | Accuracy

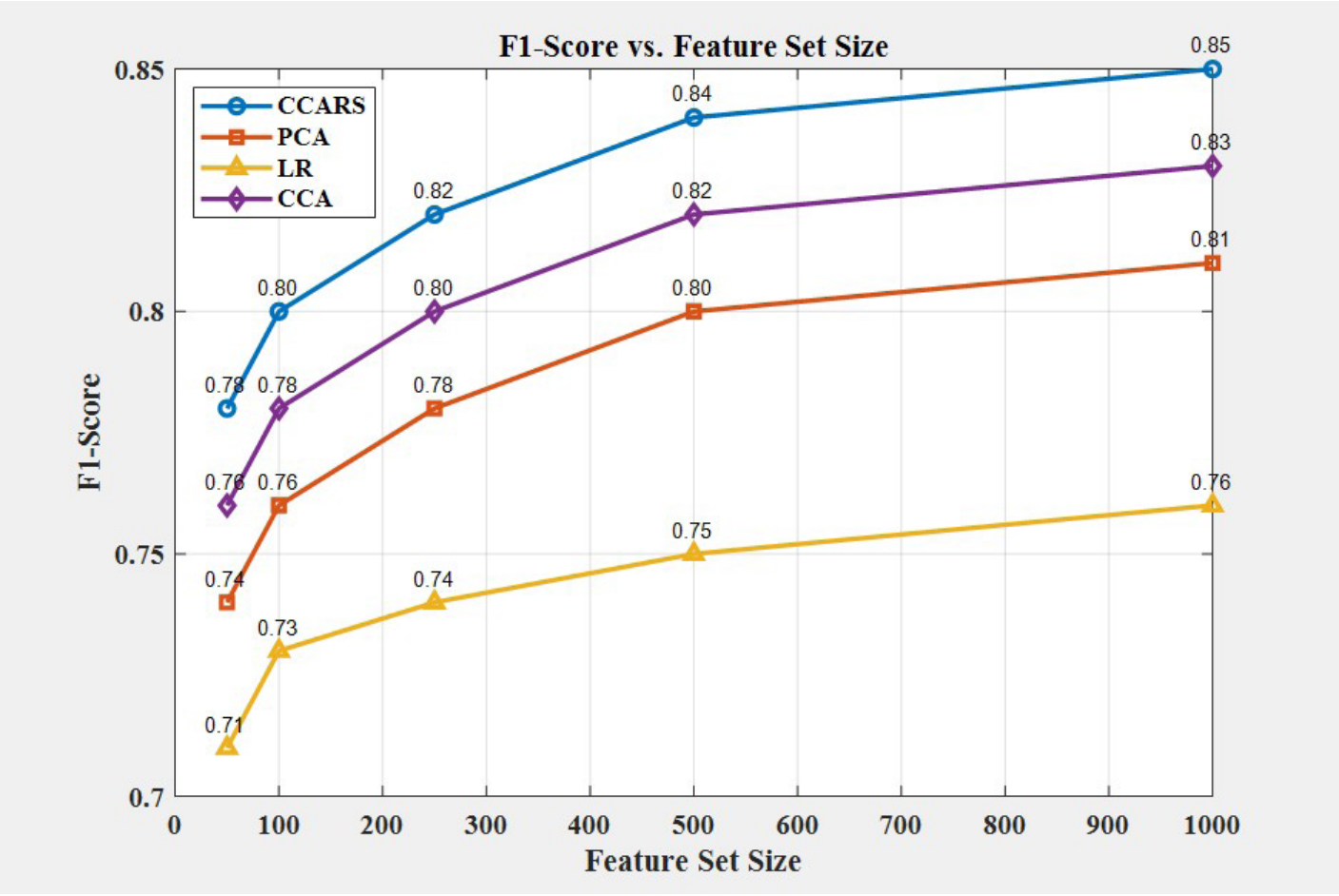

Fig 3 | F1 score

Table 5 | F1-score versus feature set size

Feature set size	rCCA-RFE selector (CCARS)	Principal component analysis	Lasso regression (LR)	Canonical correlation analysis
50	0.78	0.74	0.71	0.76
100	0.80	0.76	0.73	0.78
250	0.82	0.78	0.74	0.80
500	0.84	0.80	0.75	0.82
1000	0.85	0.81	0.76	0.83

Computational Time Analysis

The computational time results provide insights into the efficiency of each algorithm. The emulation analysis is shown in Table 5.

Table 6 | Computational time versus feature set size

Feature set size	rCCA-RFE selector (CCARS) (s)	Principal component analysis (s)	Lasso regression (LR) (s)	Canonical correlation analysis (s)
50	2.50	1.80	2.00	2.20
100	3.20	2.40	2.50	2.70
250	4.00	2.99	3.00	3.54
500	5.50	3.97	4.00	4.51
1000	7.00	4.98	5.00	6.00

The CCARS algorithm required consistently longer computational times compared to PCA, LR, and CCA across all feature set sizes (Table 6). At 50 features, CCARS took 0.7 s longer than PCA, 0.5 s longer than LR, and 0.3 s longer than CCA. For 100 features, it required 0.8 s more than PCA, 0.7 s more than LR, and 0.5 s more than CCA. At 250 features, CCARS added 1.0 s over both PCA and LR, and 0.5 s over CCA. With 500 features, CCARS took 1.5 s more than PCA and LR, and 1.0 s more than CCA.

The present study demonstrates that while CCARS takes a longer time to execute than other algorithms because of the multiple feature selection process, it still yields better results in terms of accuracy and F1-score, as shown in Figures 2 and 3. The simulation results reveal that CCARS has a much higher potential than the four methods for integrating multi-omics data in leukemia subtype/risk prediction successfully.

The simulated analysis confirms that on the issues of accuracy and F1-score, the CCARS algorithm surpasses PCA, LR, and CCA and proves its efficiency to integrate multi-omics data for early-leukemia detection. That is why classification with the help of the function is slower, but, nevertheless, it should be used because of significant improvements in classification results.

DOI: https://doi.org/10.70389/PJS.100274 | Premier Journal of Science 2026;21:100274

Table 7 | Top selected features across omics layers

Omics layer	Feature (Gene/CpG/CNV)	Selection frequency (%)	Importance score
RNA-seq	FLT3	92	0.85
	NPM1	89	0.82
DNA methylation	cg123456	87	0.78
	cg789012	85	0.75
CNV	chr8_gain	90	0.80
	chr7_loss	88	0.77

Biological Interpretation and Feature Analysis

Top features of every omics layer were examined on the basis of the selection frequency across folds to enhance interpretability (Table 7). Orthologous genes like FLT3, NPM1, and DNMT3A were repeatedly chosen, which shows that the model was stable. The pathway enrichment analysis showed the associations with hematopoiesis, cell cycle regulation, and leukemia progression. The results are consistent with established clinical biomarkers, indicating that the put forward CCARS framework embodies both clinically relevant and biologically meaningful trends with respect to detecting leukemia.

The stability of feature selection was determined by computing the frequency of the selected features in the outer cross-validation folds. Attributes with a high frequency of selection are used to show the strength of the CCARS framework and its stability. The findings prove that some salient features have been repeatedly chosen by folds, which indicates the consistency of the selected approach.

To explore further on biological relevance, the biological significance of the chosen gene set was analyzed through standard enrichment tools on pathway enrichment analysis. The analysis provided evidence of considerable enrichment in the pathways associated with hematopoiesis, cell proliferation, and leukemia progression, which also indicates the biological validity of the chosen properties.

Some of the highest-ranked genes of the model, including FLT3 and NPM1, are familiar biomarkers linked to acute myeloid leukemia and have been extensively featured in the literature. The fact that they are found in the chosen set of features shows that the offered framework manages to capture the clinically relevant signals.

Conclusion

The paper introduced a new machine learning model named CCARS to identify early signs of leukemia on the basis of multi-omics data. The proposed framework combines CCA and RFE in order to determine the most relevant features for each of the omics layers. It has been experimentally proven that the CCARS framework delivers a better classification performance over the current methods of PCA, lasso regression, and CCA especially in accuracy and F1-score. The small number of patients with 173 AML that were used in the experiments is one of the limitations of the current work. Further studies will be conducted in the future to prove the framework presented with the help of larger multi-cohort statistics that will help determine its feasibility and solidity. One of the limitations of this study is that there was no external multi-omics dataset of the same modalities. The external validation was performed with the help of gene expression data alone, which measures a partial generalization of the framework. Further research will be done in the future, whereby the suggested framework will be tested on multi-omics datasets that are fully matched to further determine whether the framework is generalizable or not.

References

1 Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience. 2022;25:103798. https://doi.org/10.1016/j.isci.2022.103798
2 Samarkhazan HS. Integrating multi-omics approaches in acute myeloid leukemia: a review. Clin. Exp. Med. 2025;25(1):311.
3 Song Y, Wang Z, Zhang G, Hou J, Liu K, Wei S, et al. Classification of acute myeloid leukemia based on multi-omics data and prognosis prediction. Mol Oncol. 2025;19(6):1836–54.
4 Li J, Wang S. Integrative analysis of epigenetic subtypes in acute myeloid leukemia: a multi-center study combining machine learning for prognostic and therapeutic insights. PLoS One. 2025;20(5):e0324380.
5 Zhang H, Zhang J. Comprehensive omics profiling in acute myeloid leukemia: a review. Curr. Proteomics. 2025;22(3):100028.
6 Hernández-Lemus E, Ochoa S. Methods for multi-omic data integration in cancer research. Front Genet. 2024;15:1425456. https://doi.org/10.3389/fgene.2024.1425456
7 Morabito A, De Simone G, Pastorelli R, Brunelli L, Ferrario M. Algorithms and tools for data-driven omics integration to achieve multilayer biological insights: a narrative review. J Transl Med. 2025;23(1):425.
8 Vahabi N, Michailidis G. Unsupervised multi-omics data integration methods: a comprehensive review. Front Genet. 2022;13:854752.
9 Ballard JL, Wang Z, Li W, Shen L, Long Q. Deep learning-based approaches for multi-omics data integration and analysis. BioData Min. 2024;17(1):38. https://doi.org/10.1186/s13040-024-00391-z
10 Alhamrani SQ, Ball GR, El-Sherif AA, Ahmed S, Mousa NO, Alghorayed SA, et al. Machine learning for multi-omics characterization of acute leukemia. Cells. 2025;14(17):1385.
11 Afroz S, Islam N, Habib MA, Reza MS, Ashad Alam M. Multi-omics data integration and drug screening in acute myeloid leukemia. Comput Biol Med. 2024;226:138-150.
12 Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46. https://doi.org/10.1016/j.csbj.2021.06.030
13 Baião ARF, Cai Z, Poulos RC, Robinson PJ, Reddel RR, Zhong Q, et al. A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. Brief. Bioinform. 2025;26(4):bbaf355.
14 Dong Y, Liao HP, Huang F, Bao YS, Guo W, Tan Z. Machine learning reveals methylation signatures associated with pediatric acute myeloid leukemia recurrence. Sci Rep. 2025;15(1):15815.
15 Alshedah N. Comparative evaluation of DIABLO and NOLAS for multi-omics integration. 2025.
16 Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2020;16(6):e1006519.
17 Argelaguet R, et al. Multi-omics factor analysis (MOFA) for unsupervised integration of multi-omics datasets. Mol Syst Biol. 2020;16(5):e9798.

18 Zeng T, et al. Similarity network fusion-based integrative clustering in cancer multi-omics data. IEEE/ACM Trans Comput Biol Bioinf. 2020;17(2):687–701.
19 Zhang AW, et al. Probabilistic modeling using iClusterPlus for multi-omics cancer subtype discovery. Bioinformatics. 2020;36(28):717–25.
20 Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet. 2020;11:554.
21 Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16(2):85–97. https://doi.org/10.1038/nrg3868
22 Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17(Suppl 2):15. https://doi.org/10.1186/s12859-015-0857-9
23 Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018;19(5):299–310. https://doi.org/10.1038/nrg.2018.4
24 Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: tools, advances and future approaches. J Mol Endocrinol. 2019;62(1):R21–45. https://doi.org/10.1530/JME-18-0055