LAUTECH Journal of Engineering and Technology 19 (4) 2025: 79-90

10.36108/laujet/5202.91.0480

Comparative Analysis of SVM and Logistic Regression for Classifying Diagnostic MicroRNA Signatures in Colorectal Cancer

[1]Mabayoje M. A., [2]Bello J., [3]Oladele T. O., [4]Akinrotimi A. O. and [5]Mabayoje A. A.

[1,3]Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Kwara State, Nigeria.

[2,]Department of Mathematical and Computer Sciences, College of Natural and Applied Sciences, Fountain University, Osogbo, Osun State, Nigeria.

[4]Department of Information Systems and Technology, Faculty of Science, King's University, Ode-Omu, Osun State, Nigeria.

[5]European University of Lefke, Gemikonagi, Lefke MERSIN 10 KKTC, Türkiye.

Article Info

Article history:

Received: June 30, 2025
Revised: Sept 03, 2025
Accepted: Sept 04, 2025

Keywords:

Genetic Algorithm, Independent Component Analysis, Colorectal Cancer, microRNA, Support Vector Machine.

Corresponding Author:
mabayoje.ma@unilorin.edu.ng

ABSTRACT

The Early and accurate classification of gene signatures is critical for improving colorectal cancer (CRC) diagnosis. While previous studies have applied machine learning to microRNA datasets, few have combined feature selection and extraction methods in a unified diagnostic pipeline. This study proposes a novel integration of Genetic Algorithm (GA) and Independent Component Analysis (ICA) for selecting and extracting relevant features from high-dimensional microRNA data. GA is used as a wrapper-based feature selection method to reduce the original 2457 features to 52, while ICA further transforms these into 12 uncorrelated components. These components are then classified using Support Vector Machine (SVM) and Logistic Regression (LR) models. Using the GA–ICA–SVM pipeline, we achieved an AUC of 0.8347, outperforming the LR model, which achieved an AUC of 0.7318. This approach demonstrates improved performance and efficiency in detecting CRC-related biomarkers and offers a reproducible framework for biomarker-based cancer diagnosis.

INTRODUCTION

Machine learning is a subset of Artificial Intelligence (AI), which continuously learns from various examples and is applied to real-world problems. Classification is a Machine Learning activity that assigns a label value to a given class and then evaluates whether a specific type belongs to that class. A simple example is the email spam filtration system, which allows users to designate emails as either “spam” or “not spam”. There are various classification issues to be encountered, and unique approaches can be employed for each difficulty (Mitsala *et al.*, 2021). Classification is a term used to describe any situation in which a specified class label must be predicted from a given data field. A training dataset is necessary for each model, including numerous inputs and outputs from which the model will learn. For the model to be trained successfully, the training data must include all conceivable issue scenarios and adequate data for each label. Because class labels are frequently returned as text values, they must be converted into an integer, such as 0 for "spam" and 1 for "no-spam." (Hameed *et al.*, 2017).

Classification accuracy helps assess a model's performance based on multiple anticipated class labels. Although classification accuracy is not the most important criterion, it is a good starting point for most classification problems. Some models may provide us with a likelihood of class membership of a specific input instead of a class label. The ROC curve may be a valuable measure of a model's accuracy in such circumstances. Classification is a core task in machine learning where input data is assigned to predefined categories or classes. A familiar example is spam detection, where emails are labeled as either "spam" or "not spam." For a model to learn accurately, it must be trained on a dataset containing representative examples for each class. Often, class labels are textual (e.g., "positive," "negative") and are typically converted to numerical values during the preprocessing stage. In machine learning, classification problems are commonly grouped into four types: binary classification, multi-class classification, multi-label classification, and imbalanced classification.

Choosing the right model often depends on which of these task types is being addressed. While accuracy is a common performance metric, other evaluation methods, such as ROC curves, are used when class distributions are uneven or when probabilistic outputs are required. Due to the numerous correlations and redundancies among human genes, several computational approaches have failed to extract a limited selection of ascribed genes in high-dimensional datasets. Studies in cancer informatics have demonstrated that data mining and machine learning are valuable tools in predicting diseases, particularly in identifying the associated genes that cause cancer (Akinrotimi and Oladele, 2018). Machine learning has been shown to perform well in cancer classification; nevertheless, it still has to be improved and made more resilient in terms of efficiency and computational cost, especially when dealing with large datasets. High-dimensional datasets contain several redundant and variant gene expressions, which reduces the accuracy and efficiency of computational algorithms used to extract the most attributable genes (Hameed *et al*., 2017).

Biological differences associated with studies or gene modifications typically cause noise in gene expression levels (Hameed *et al*., 2021). As a result, finding the ascribed genes in high-dimensional datasets is difficult unless a rigorous analysis and selection method is applied.

MicroRNAs can be referred to as oncomiRs and tumor suppressor microRNAs, as their expression patterns have been demonstrated to be different among tissues and bodily fluids compared to the expected. As a result, they can be used as diagnostic, prognostic, and predictive biomarkers of CRC. As a result, identifying prognostic and predictive biomarkers is critical for certifying the purity standard in cancer genomics (Fadaka *et al*., 2018). Meanwhile, due to numerous molecular approaches, an increasing number of genes are being linked to CRC. Interferon genetic variants, particularly interferon-gamma and interferon regulatory factors, have been linked to an increased risk of CRC and a shorter survival time following diagnosis. Early identification of CRC is a significant difficulty worldwide, which means that current treatment options are being delivered late after the tumor has spread. If tumors are found early enough and polyps are surgically removed, the incidence and mortality rate of CRC may be reduced (Fadaka *et al*., 2019).

Despite growing interest in the application of machine learning for cancer diagnosis, many existing studies focus on either feature selection or dimensionality reduction in isolation. Few have attempted to integrate a wrapper-based selection method, such as the Genetic Algorithm (GA), with a transformation-based extraction method, like Independent Component Analysis (ICA), especially in the context of microRNA (miRNA) data for colorectal cancer (CRC). This gap limits the efficiency and accuracy of current diagnostic models. Therefore, this study aims to develop and evaluate a hybrid GA–ICA approach for selecting and extracting informative features from high-dimensional CRC miRNA datasets. The resulting features are then classified using Support Vector Machine (SVM) and Logistic Regression (LR) to assess and compare performance (the GA–ICA–SVM pipeline). This integrated approach is designed to improve classification accuracy and offer a reproducible framework for early CRC detection using miRNA biomarkers.

RELATED WORK

Several studies have explored the role of non-coding RNAs, particularly microRNAs (miRNAs), in colorectal cancer (CRC) development, diagnosis, and prognosis. For example, Fadaka *et al.* (2019) reported that miRNAs play a critical role in modulating the expression of oncogenes and tumor suppressor genes, making them promising biomarkers across different stages of CRC. Their deregulation influences multiple cancer-related pathways, suggesting their use in therapeutic monitoring and disease staging. Similarly, Zhi *et al.* (2018) performed a meta-analysis on five gene expression datasets to identify differentially expressed genes (DEGs) in metastatic vs. non-metastatic CRC samples. Using Support Vector Machine (SVM) classifiers and protein-protein interaction networks, they isolated genes such as CREB1, CUL7, and SSR3 as potential biomarkers for predicting metastasis. In another study, Herreros-Villanueva *et al.* (2019) evaluated a six-miRNA signature in plasma samples from 297 individuals. Their classifier achieved an AUC of 0.92, with high sensitivity and specificity, distinguishing CRC and advanced adenomas (AA) from healthy controls. This affirmed the diagnostic power of circulating miRNAs. Wang *et al.* (2020) provided a broader perspective by analyzing the current applications of artificial intelligence (AI) in CRC, highlighting its potential across diagnostics, treatment planning, and prognosis prediction.

Similarly, Di *et al.* (2020) employed a hybrid model of SVM and LASSO to integrate multiple datasets and identify miRNAs with high predictive value for CRC. Moreover, Yuan (2021) introduced a framework that combines the mRMR feature selection method with four classifiers: Random Forest, SVM, k-Nearest Neighbors (kNN), and Decision Tree (DT), to detect cancer subtypes using extracellular miRNA data. Their workflow prioritized prediction accuracy and early diagnosis through robust dimensionality reduction. However, despite these advances, few studies have integrated both a wrapper-based feature selection technique, such as the Genetic Algorithm (GA), with a transformation-based method, like Independent Component Analysis (ICA), in a unified pipeline. Existing works often employ one or the other, and many do not describe how these techniques interact during the preprocessing of high-dimensional miRNA data. Furthermore, kernel selection and cross-validation methods are often omitted or under-detailed, which limits the reproducibility and generalizability of results.

This study addresses these gaps by combining GA for optimal feature subset selection with ICA for feature extraction, followed by classification using SVM and Logistic Regression. This approach not only improves interpretability by reducing data dimensionality but also enhances classification accuracy. Our pipeline is applied to a curated CRC microRNA dataset and evaluated using multiple metrics, including AUC, sensitivity, specificity, and cross-validation, providing a transparent and reproducible methodology for biomarker-based cancer classification.

METHODOLOGY

Overview of Approach

This study follows a structured pipeline that involves data acquisition, feature selection using a Genetic Algorithm (GA), dimensionality reduction using Independent Component Analysis (ICA), and classification using Support Vector Machine (SVM) and Logistic Regression (LR). The full workflow is shown in Figure 2.

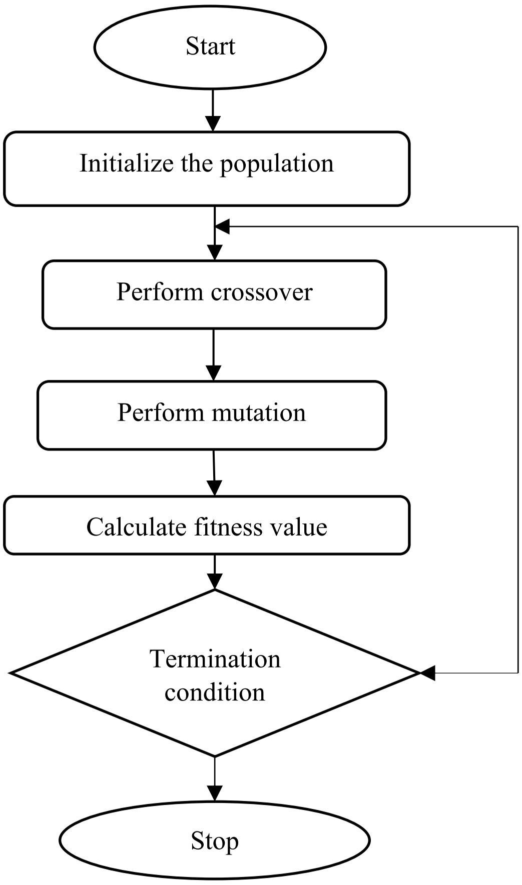


Figure 1: Flowchart Representation of Genetic Algorithm (Source: Asir *et al.*, 2016).

Genetic Algorithm for Feature Selection.

A Genetic Algorithm (GA) is employed as a wrapper-based feature selection method to identify the most informative genes from a high-dimensional microRNA dataset. The GA process begins with population initialization and iteratively performs crossover, mutation, and fitness evaluation. The goal is to retain features that best contribute to classification accuracy. In this study, a standard implementation of GA was used, without custom enhancements. The algorithm selects optimal subsets of features based on performance, and the number of selected features was reduced from 2457 to 52 after this stage. Figure 1 presents the flowchart of the GA process, and Algorithm 1 outlines the steps in pseudocode.

Algorithm 1: Genetic Algorithm

```
Necessitate. Set parameters nPop = m, tmax, t = 0;
Confirm: optimum feature subset with the maximum suitable rate.
while (t<=tmax) do
    Create pop a, tmax;
    For k = 1 to a do
        Parents [al, a2] = system selection (a, nPop)
        Child = Xor[al, a2]
        M u = mutation [Child}
    End for
    Replace a with      Child1, Child2, Childm
    t = t+1;
End while
Save the Highest fitness value;
```

a = population size, r = random number 0 to 1, chrome = certain or non-certain feature through threshold δ, set value = 0.5, and a = threshold amount of picked features. Selecting the maximum fit features from the predictive datasets is the primary challenge of the GA technique. This study involves four significant phases, and the experimental workflow is illustrated in Figure 1. In processing the experimental data, feature selection using the Genetic Algorithm was performed on a microRNA colorectal cancer dataset containing 2457 instances and 7 attributes. The loaded data is depicted in Figure 2, which displays relevant information extracted from the dataset using the procedure outlined in Figure 3. The goal of ICA is to discover uncorrelated linear modifications (latent components) of the original predictor variables that covariate firmly with the response variables. To reduce the dimensionality of the specified functions, functionality extraction provides new variables as variants. This approach leverages beneficial attributes while minimizing negative ones. It functions by replacing the initial variables (numeric) with new numeric variables; it captures the most defining feature. It functions by replacing the initial variables (numeric) with new numeric variables; it captures the most defining feature of Santos and Cunha (2015).

Independent Component Analysis (ICA) for Feature Extraction

1CA is a valued leeway of PCA with conservative layers, as it allows for the visor parting of independent bases from their linear grouping. The fact of ICA is the possession of the uncorrelation of the general PCA. Built a_n x b on data matrix P, whose rows ri (d=l…, a) reckon to observational variables and whose columns kd (d-1..., b) are the entities of the matching variables, the ICA model of P can be written as shown in equation 1:

$$P = AS \quad (1)$$

With a complete overview, A is a x a fusion matrix, where S is a a x b is a basis matrix. Independent components are the original variables stored in rows of S; that is, the variables detected are linearly composed of independent components. The independent components are achieved by learning the precise linear groupings of the experimental variables since mixing can be inverted, as shown in equation 2:

$$U = S = A\text{-}1P = WP \quad (2)$$

After GA-based feature selection, Independent Component Analysis (ICA) is applied to further reduce redundancy and isolate statistically independent components. ICA transforms the selected features into a new space where they are uncorrelated and capture the most meaningful variation for classification tasks.

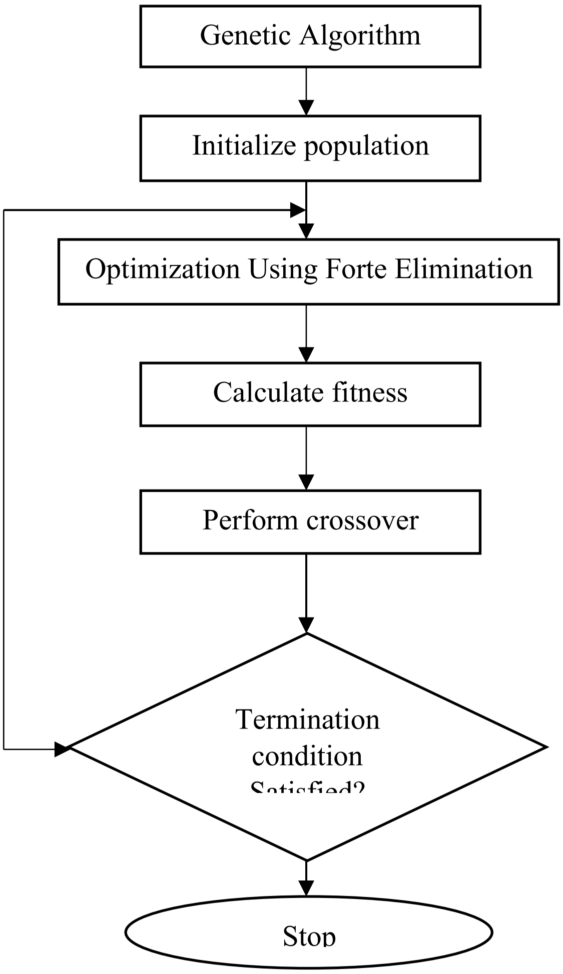


Figure 2: Flowchart for the proposed Feature Selection Technique

From the 52 features selected by GA, ICA extracted 12 independent components, effectively reducing the data from 2457 to 12 dimensions. This helps eliminate noise and improve classifier performance. The ICA transformation process is illustrated in Figure 5.

Classification Models

As the number of features in a dataset increases, so does the cost of computation. To reduce the computational cost, the number of features is typically reduced (Omololu and Abolore, 2018). As such, the genetic algorithm was used in the feature selection phase of this study to select the optimal subset of relevant genes from the original data without halting the dimensional space. The feature extraction technique uses ICA on the selected features from the first phase to the second phase. The reduced dataset (12 ICA features) was used as input for two classifiers: Support Vector Machine (SVM) and Logistic Regression (LR). SVM was evaluated with linear, polynomial, and RBF kernels, while LR served as a baseline comparator. The models were trained using both a training/test split and 5-fold cross-validation. Classification performance was measured using AUC, sensitivity, specificity, PPV, and NPV. The entire model training pipeline is shown in Figure 3.

Model Training Setup

To evaluate the classification performance of the proposed approach, Support Vector Machine (SVM) and Logistic Regression (LR) models were implemented using the R programming environment. After the feature selection and extraction phases using Genetic Algorithm (GA) and Independent Component Analysis (ICA), the reduced dataset was used for training and testing both classifiers. For the SVM classifier, two kernel functions were explored: the Gaussian Radial Basis Function (RBF) and the Polynomial kernel. Model training was performed with default SVM parameters where applicable, and the kernel parameters (e.g., gamma and degree) were selected based on empirical performance during preliminary runs. Similarly, the LR model was trained using a standard implementation without regularization, assuming a binary logistic regression model structure. A 5-fold cross-validation scheme was adopted to assess the models' generalization performance. The dataset was randomly divided into five equal subsets: in each iteration, four folds were used for training, and the remaining one was used for testing. This process was repeated five times to ensure that each subset served as the test set once. Additionally, the models were trained on the full training set and evaluated separately on a held-out test set, providing an independent assessment of classification performance. Key metrics reported include sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the Area Under the ROC Curve (AUC). To ensure reproducibility, the random seed was fixed during data splitting and model training phases.

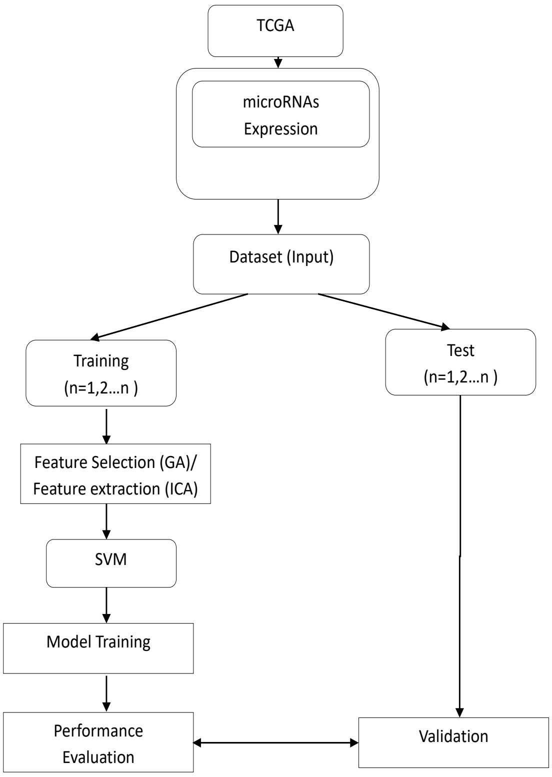


Figure 3: Workflow Diagram Depicting the Mechanism of the Proposed Colorectal Cancer Diagnosis Model

Model Settings and Implementation Details

The classification models used in this study, namely Support Vector Machine (SVM) and Logistic Regression (LR), were implemented using the R programming environment, specifically within the RStudio IDE. The primary packages used included e1071 for SVM and glm for Logistic Regression.

For the SVM classifier, three kernel types were evaluated: (a) Linear Kernel (b) Polynomial Kernel (degree = 3) (c) Radial Basis Function (RBF). The Gaussian RBF kernel showed the best performance in terms of AUC and overall classification accuracy. Where applicable, the following hyperparameters were configured:

(a) Cost (C): Set to 1 (default)

(b) Gamma (γ): Automatically computed as 1/n features for RBF

(c) Degree: Set to 3 for the Polynomial kernel

(d) Tolerance: 0.001

(e) Cross-validation folds

(h) The Logistic Regression model was used in its standard binary form without regularization. It served as a baseline comparator for the SVM classifier. To ensure reproducibility, a fixed random seed (e.g., set. seed (42)) was used during data splitting, feature selection, and model training stages. All experiments were conducted on a standard desktop system with R version 4.0.3 or later.

Description of Datasets

The colorectal cancer dataset used in this study was obtained from kaggle.com. This dataset contains cases that are either not harmful or harmful. Using an in silico technique, five potential microRNAs, decoded as miR-1 to 5, were identified, and their target genes were selected using three separate target prediction tools (TargetScan, miRDB, and miRDIP) to generate seven target genes: APC, GNAS, EGFR, TCF7L2, KRAS, IGF1R, and CASP8. In this study, the functional determination was based on the sequences of these microRNAs and the promoter sequences of their targets.

Performance Evaluation Metrics

The models' predictive performance is examined using the Cross-Validation approach to estimate how each model performs outside the sample in a new dataset, also known as test data. When data fit into a model, cross-validation procedures are used to fit it to a training dataset. The dataset only contains information on how the models perform with training data if cross-validation is not used.

In an ideal world, new data would be used to assess the models' performance in terms of prediction accuracy. Theories in science are evaluated based on their ability to predict future outcomes. It's a popular methodology since it's straightforward to grasp and produces a less biased or optimistic assessment of model competence than other approaches, such as a fundamental train/test split. The approach features a parameter called k that specifies how many groups a given data sample should be divided into. As a result, the process is frequently referred to as k-fold cross-validation. It’s a great way to test and quantify the accuracy of classifiers, as it divides the training set into k subsets at random, with one of the k subsets used for testing and the rest for training. A 5-fold cross-validation method was used. It splits the dataset into 5 parts, trains on 4, and tests on 1, repeating for all combinations of train-test splits. This technique helps avoid overfitting of the training set, especially in small datasets with many attributes. The receiver operating characteristic (ROC) curve's area under the curve (AUC) and other cross-validation data (sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV] to create summary performance estimates.

Performance Evaluation Metrics

The expected and tangible outcomes were used to create ROC curves. The AUCs for the test datasets were calculated and compared to assess how well the models discriminated. P-values were calculated using DeLong's approach to compare AUCs based on SVM and MLR models (Yang et al, 2023). When the cutoff value in the SVM model was set to the default value (0), the following formulae were used to determine sensitivity, specificity, PPV, and NPV.

$$\text{Sensitivity} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FN} \quad (4)$$

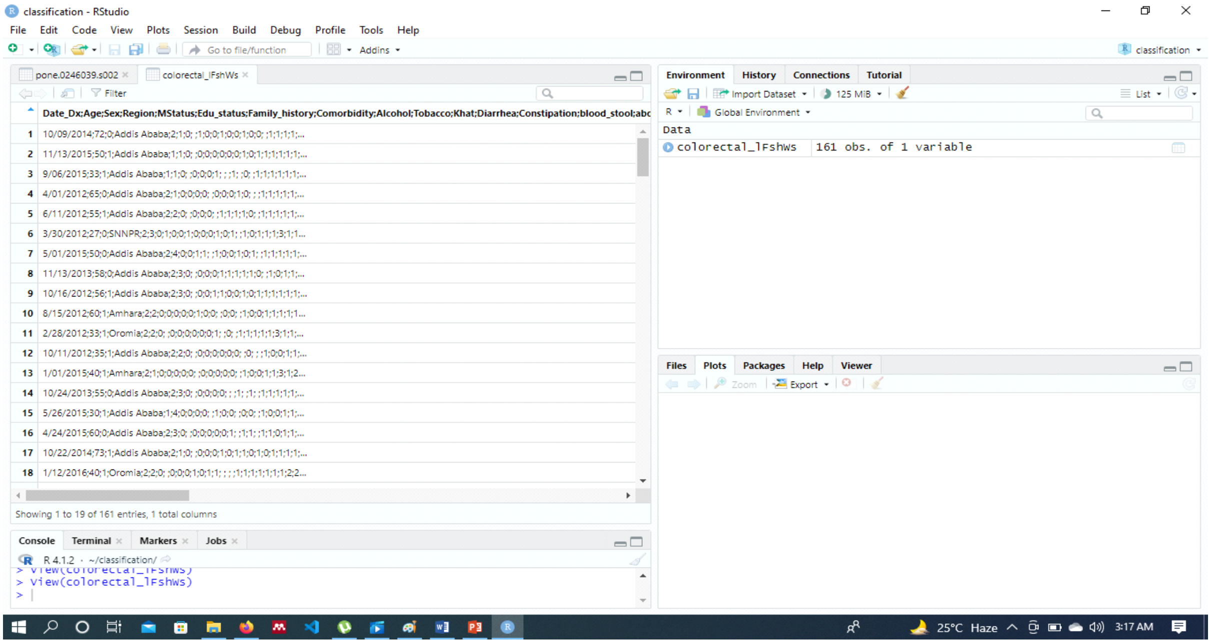
$$\text{PPV} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{NPV} = \frac{TN}{TN+FN} \quad (6)$$

Where: TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives.

RESULTS AND DISCUSSION

Feature selection and extraction methods were created on the R programming platform, followed by classification approaches. The findings of the investigations for the suggested model are presented in this section. These approaches were used with the help of an improved genetic algorithm (GA) and feature extraction (ICA). This work uses a dimensionality reduction strategy with SVM classification algorithms on a colorectal cancer dataset. After applying Genetic Algorithm (GA) for feature selection and Independent Component Analysis (ICA) for dimensionality reduction, the final feature set was used to train both Support Vector Machine (SVM) and Logistic Regression (LR) models. For SVM, three kernel types (linear, polynomial, and radial basis function) were evaluated, while LR was implemented as a baseline linear classifier. Each model was trained using a 5-fold cross-validation strategy and tested on a separate validation set to assess generalization. Performance metrics such as AUC, sensitivity, specificity, PPV, and NPV were computed and compared across both models. There are 7 characteristics and 2457 gene expression levels in the data, standardized. The integrated development region on R Studio that was utilized for the model is shown in Figure 4.


classification - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
classification
pone.0246039.s002
colorectal_lFshWs
Filter
Environment History Connections Tutorial
Import Dataset
125 MiB
List
Global Environment
Files Plots Packages Help Viewer
Zoom
Export
Console Terminal Markers Jobs
R 4.1.2 · ~/classification/
25°C Haze
3:17 AM

	Date_Dx;Age;Sex;Region;MStatus;Edu_status;Family_history;Comorbidity;Alcohol;Tobacco;Khat;Diarrhea;Constipation;blood_stool;abc
1	10/09/2014;72;0;Addis Ababa;2;1;0; ;1;0;0;1;0;0;1;0;0; ;1;1;1;1;...
2	11/13/2015;50;1;Addis Ababa;1;1;0; ;0;0;0;0;0;0;1;0;1;1;1;1;1;1;...
3	9/06/2015;33;1;Addis Ababa;1;1;0; ;0;0;0;1; ; ;1; ;0; ;1;1;1;1;1;1;...
4	4/01/2012;65;0;Addis Ababa;2;1;0;0;0;0; ;0;0;0;1;0; ; ;1;1;1;1;1;...
5	6/11/2012;55;1;Addis Ababa;2;2;0; ;0;0;0; ;1;1;1;1;0; ;1;1;1;1;1;...
6	3/30/2012;27;0;SNNPR;2;3;0;1;0;0;1;0;0;0;1;0;1; ;1;0;1;1;1;3;1;1...
7	5/01/2015;50;0;Addis Ababa;2;4;0;0;1;1; ;1;0;0;1;0;1; ;1;1;1;1;1;...
8	11/13/2013;58;0;Addis Ababa;2;3;0; ;0;0;0;1;1;1;1;1;0; ;1;0;1;1;...
9	10/16/2012;56;1;Addis Ababa;2;3;0; ;0;0;1;1;0;0;1;0;1;1;1;1;1;1;...
10	8/15/2012;60;1;Amhara;2;2;0;0;0;0;0;1;0;0; ;0;0; ;1;0;0;1;1;1;1;1...
11	2/28/2012;33;1;Oromia;2;2;0; ;0;0;0;0;0;0;1; ;0; ;1;1;1;1;1;3;1;1;...
12	10/11/2012;35;1;Addis Ababa;2;2;0; ;0;0;0;0;0;0; ;0; ; ;1;0;0;1;1;...
13	1/01/2015;40;1;Amhara;2;1;0;0;0;0;0; ;0;0;0;0;0; ;1;0;0;1;1;3;1;2...
14	10/24/2013;55;0;Addis Ababa;2;3;0; ;0;0;0;0; ; ;1; ;1; ;1;1;1;1;1;...
15	5/26/2015;30;1;Addis Ababa;1;4;0;0;0;0; ;1;0;0; ;0;0; ;1;0;0;1;1;...
16	4/24/2015;60;0;Addis Ababa;2;3;0; ;0;0;0;0;0;1; ;1;1; ;1;1;0;1;1;...
17	10/22/2014;73;1;Addis Ababa;2;1;0; ;0;0;0;1;0;1;1;0;1;0;1;1;1;1;...
18	1/12/2016;40;1;Oromia;2;2;0; ;0;0;0;1;0;1;1; ; ; ;1;1;1;1;1;1;1;2;2...

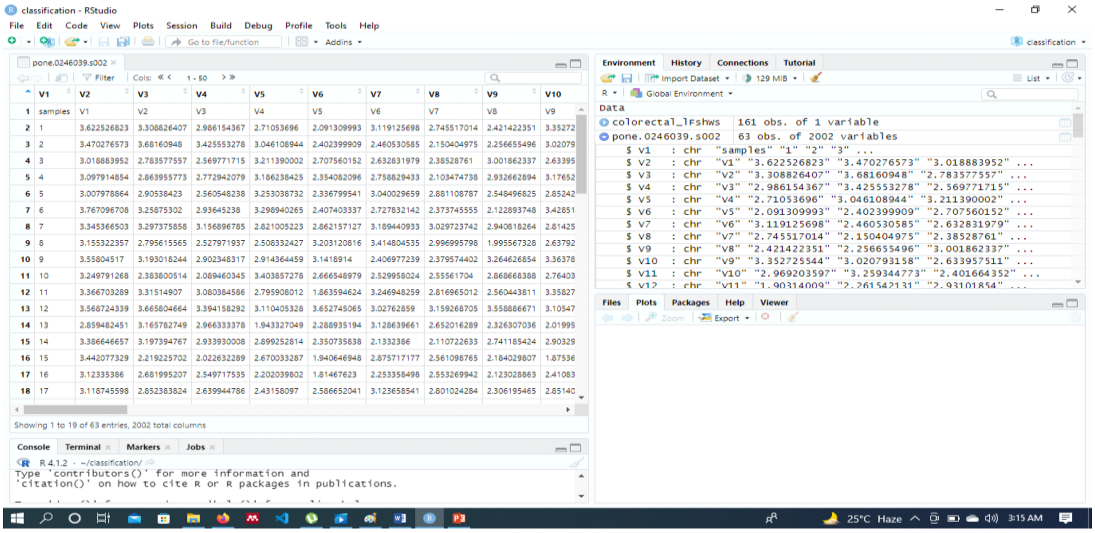
Showing 1 to 19 of 161 entries, 1 total columns

```
Data
colorectal_lFshWs    161 obs. of 1 variable
```

```
> View(colorectal_lFshWs)
> View(colorectal_lFshWs)
> |
```

Figure 4: R programming environment showing execution of the proposed GA–ICA–SVM model on the colorectal cancer microRNA dataset

This study looks at microRNAs in colorectal cancer data and genes that are sensitive and resistant to the disease; therefore, to minimize the curse of dimensionality, the ICA technique, which is a non-linear approach, was used for the data processing stage. Figure 5 depicts the ICA procedure. ICA finds and eliminates uncorrelated attributes (variables) to determine maximum variance with fewer latent



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	samples	V1	V2	V3	V4	V5	V6	V7	V8	V9
2	1	3.622526823	3.308826407	2.986154367	2.71053696	2.091309993	3.119125698	2.745517014	2.421422351	3.35272
3	2	3.470276573	3.68160948	3.425553278	3.046108944	2.402399909	2.460530585	2.150404975	2.256655496	3.02079
4	3	3.018883952	2.783577557	2.569771715	3.211390002	2.707560152	2.632831979	2.38528761	3.001862337	2.63395
5	4	3.097914854	2.863955773	2.772942079	3.186238425	2.354082096	2.758829433	2.103474738	2.932662894	3.17652
6	5	3.007978864	2.90538423	2.560548238	3.253038732	2.336799541	3.040029659	2.881108787	2.548496825	2.85242
7	6	3.767096708	3.25875302	2.93645238	3.298940265	2.407403337	2.727832142	2.373745555	2.122893748	3.42851
8	7	3.345366503	3.297375858	3.156896785	2.821005223	2.862157127	3.189440933	3.029723742	2.940818264	2.81425
9	8	3.155322357	2.795615565	2.527971937	2.508332427	3.203120816	3.414804535	2.996995798	1.995567328	2.63792
10	9	3.55804517	3.193018244	2.902348317	2.914364459	3.1418914	2.406977239	2.379574402	3.264626854	3.36378
11	10	3.249791268	2.383800514	2.089460345	3.403857278	2.666548979	2.529958024	2.55561704	2.868668388	2.76403
12	11	3.366703289	3.31514907	3.080384586	2.795908012	1.863594624	3.246948259	2.816965012	2.560443811	3.35827
13	12	3.568724339	3.665804664	3.394158292	3.110405328	3.652745065	3.02762859	3.159268705	3.558886671	3.10547
14	13	2.859482451	3.165782749	2.966333378	1.943327049	2.288935194	3.128639661	2.652016289	2.326307036	2.01995
15	14	3.386646657	3.197394767	2.933930008	2.899252814	2.350735838	2.1332386	2.110722633	2.741185424	2.90329
16	15	3.442077329	2.219225702	2.022632289	2.670033287	1.940646948	2.875717177	2.561098765	2.184029807	1.87536
17	16	3.12335386	2.681995207	2.549717535	2.202039802	1.81467623	2.253358498	2.553269942	2.123028863	2.41083
18	17	3.118745598	2.852383824	2.639944786	2.43158097	2.586652041	3.123658541	2.801024284	2.306195465	2.85140

Showing 1 to 19 of 63 entries, 2002 total columns

```
Data
colorectal_lFshws    161 obs. of 1 variable
pone.0246039.s002    63 obs. of 2002 variables
   $ V1   : chr  "samples" "1" "2" "3" ...
   $ V2   : chr  "V1" "3.622526823" "3.470276573" "3.018883952" ...
   $ V3   : chr  "V2" "3.308826407" "3.68160948" "2.783577557" ...
   $ V4   : chr  "V3" "2.986154367" "3.425553278" "2.569771715" ...
   $ V5   : chr  "V4" "2.71053696" "3.046108944" "3.211390002" ...
   $ V6   : chr  "V5" "2.091309993" "2.402399909" "2.707560152" ...
   $ V7   : chr  "V6" "3.119125698" "2.460530585" "2.632831979" ...
   $ V8   : chr  "V7" "2.745517014" "2.150404975" "2.38528761" ...
   $ V9   : chr  "V8" "2.421422351" "2.256655496" "3.001862337" ...
   $ V10  : chr  "V9" "3.352725544" "3.020793158" "2.633957511" ...
   $ V11  : chr  "V10" "2.969203597" "3.259344773" "2.401664352" ...
   $ V12  : chr  "V11" "1.90314009" "2.261542131" "2.93101854" ...
```

```
R 4.1.2 · ~/classification/
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
```

Figure 5: ICA feature extraction process applied to the reduced microRNA dataset after GA-based feature selection.

Table 1: Results for the Metrics Used in Classification Schemes I And II.

Model	Dataset	Sensitivity	Specificity	PPV	NPV	AUC
Classification	Test	0.7715	0.7503	0.4926	0.9127	0.8347
(Using SVM)	Training	0.7938	0.7169	0.4550	0.9211	0.8383
	5-fold cross-validation	0.7765	0.7027	0.4388	0.9130	0.8242
Classification	Test	0.7359	0.6254	0.5061	0.8195	0.7318
(Using LR)	Training	0.7092	0.6590	0.6729	0.8087	0.7393
	5-fold cross-validation	0.7059	0.6589	0.5293	0.8054	0.7357

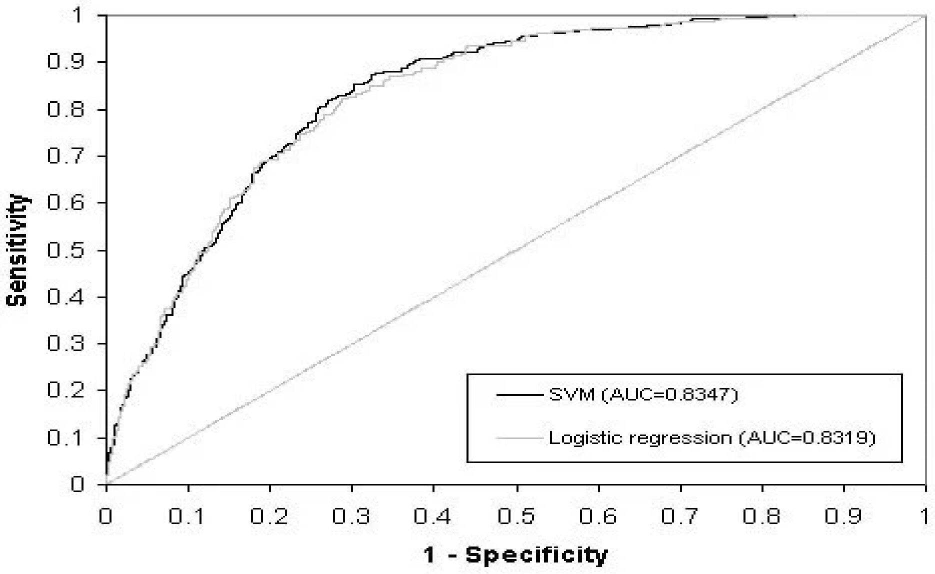


Figure 6: ROC curves for Classifications with SVM and logistic regression models.

components. In Figure 5, the output components represent the most informative, uncorrelated features.

In this study, ICA is used to reduce the dimensionality of the data and provide crucial gene information that may be used for further research. SVM-Gaussian kernel and Polynomial kernel are used in the classification algorithm, implemented using the R platform. After applying Genetic Algorithm (GA) for feature selection and Independent Component Analysis (ICA) for dimensionality reduction, the final feature set was used to train both Support Vector Machine (SVM) and Logistic Regression (LR) models. For SVM, three kernel types (linear, polynomial, and radial basis function) were evaluated, while LR was implemented as a baseline linear classifier. Each model was trained using a 5-fold cross-validation strategy and tested on a separate validation set to assess generalization. Performance metrics such as AUC, sensitivity, specificity, PPV, and NPV were computed and compared across both models.

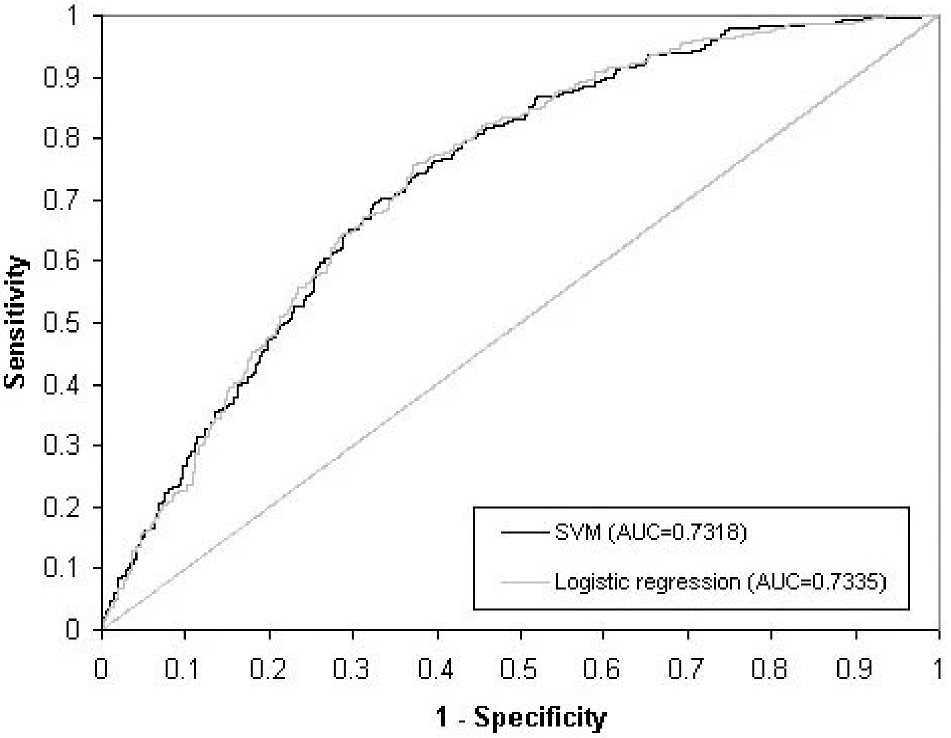


Figure 7. ROC curves for Classifications with SVM and logistic regression models.

Table 2: Best performance

Model	**The Area under the Curve**		
	Linear	**Polynomial**	**Radial basis function**
Classification (Using SVM)	0.8332	0.7655	0.8347*
Classification (Using LR)	0.7318*	0.6673	0.7259

Summary

From Tables 1 and 2, it can be observed that the highest classification rate is 0.8347, achieved using the Linear, Polynomial, and Radial Bias Function, while the highest classification rate is 0.7318, also achieved using the Linear, Polynomial, and Radial Bias Function. Experiments were conducted using a dimensionality reduction approach that employed an improved genetic algorithm with feature extraction (ICA) for microRNA classification in the colorectal cancer dataset. The classifier employs an SVM method, and the accuracy of the GA-0 + ICA performance metrics is approximately 83%, while that of Linear Regression is approximately 73%.

CONCLUSION

This study analyzed a high-dimensional microRNA dataset for colorectal cancer classification using a hybrid feature engineering and classification pipeline. Genetic Algorithm (GA) was used for selecting the most relevant features, followed by Independent Component Analysis (ICA) to extract uncorrelated components. These were then classified using Support Vector Machine (SVM) and Logistic Regression (LR) models. Key findings from this study include: (a) The GA–ICA–SVM approach achieved a higher classification accuracy (AUC = 0.8347) than Logistic Regression (AUC = 0.7318). (b) Dimensionality was successfully reduced from 2457 features to 12 without sacrificing predictive performance. (c) The use of both wrapper-based selection (GA) and transformation-based extraction (ICA) improved model efficiency and interpretability. This study used a single dataset with limited sample diversity. Further validation on larger and independent datasets is needed to confirm the generalizability of the findings. Future work could also explore integrating other classifiers or optimizing GA parameters for even better performance.

REFERENCES

Akinrotimi, A., and Oladele, R. (2018). Predictive analysis in health data using a back propagation neural network (BPNN) and C4.5 decision tree. Advances in Multidisciplinary and Scientific Research Journal, 4(2), 47–54.

Di, Z., Di, M., Fu, W., Tang, Q., Liu, Y., Lei, P., Gu, X., Liu, T. and Sun, M. (2020). Integrated analysis identifies a nine-microRNA signature biomarker for diagnosis and prognosis in colorectal cancer. Frontiers in Genetics, 11, 192. https://doi.org/10.3389/fgene.2020.00192

Fadaka, A. O., Klein, A., and Pretorius, A. (2019). In silico identification of microRNAs as candidate colorectal cancer biomarkers. Tumor Biology, 41(11). https://doi.org/10.1177/1010428319883721

Fadaka, A. O., Ojo, B. A., Adewale, O. B., Esho, T., and Pretorius, A. (2018). Effect of dietary components on miRNA and colorectal carcinogenesis. Cancer Cell International, 18, 1–14. https://doi.org/10.1186/s12935-018-0631-y

Fadaka, A. O., Pretorius, A., and Klein, A. (2019). Biomarkers for stratification in colorectal cancer: MicroRNAs. Cancer Control, 26(1). https://doi.org/10.1177/1073274819862784

Hameed, S. S., Hassan, R., and Muhammad, F. F. (2017). Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and AGBPSO-SVM algorithm. PLOS ONE, 12(11), e0187371. https://doi.org/10.1371/journal.pone.0187371

Hameed, S. S., Hassan, R., Hassan, W. H., Muhammadsharif, F. F., and Latiff, L. A. (2021). HDG-Select: A novel GUI-based application for gene selection and classification in high-dimensional datasets. PLOS ONE, 16(1), e0246039. https://doi.org/10.1371/journal.pone.0246039

Herreros-Villanueva, M., Duran-Sanchon, S., Martín, A. C., Pérez-Palacios, R., Vila-Navarro, E., Marcuello, M., Díaz-Centeno, M., Cubiella, J., Diez, M. S., Bujanda, L., Lanas, Á., Jover, R., Hernández, V., Quintero, E., Lozano, J. J., García-Cougil, M., Martínez-Arranz, I., Castells, A., Gironella, M. and Arroyo, R. (2019). Plasma microRNA signature validation for early detection of colorectal cancer. Clinical and Translational Gastroenterology, 10(1). https://doi.org/10.14309/ctg.0000000000000000

Mitsala, A., Tsalikidis, C., Pitiakoudis, M., Simopoulos, C., and Tsaroucha, A. K. (2021). Artificial intelligence in colorectal cancer screening, diagnosis and treatment: A new era. Current Oncology, 28(3), 1581–1607. https://doi.org/10.3390/curroncol28030149

Omololu, A. A., and Abolore, M. M. (2018). Comparative evaluation of filter and wrapper-based approaches for microarray cancer classification. World Journal of Research and Review, 7(4), 25-27

Santos, P., and Cunha, T. M. (2015). Uterine sarcomas: Clinical presentation and MRI features. Diagnostic and Interventional Radiology, 21(1), 4–10. https://doi.org/10.5152/dir.2014.14244

Soufan, O., Kleftogiannis, D., Kalnis, P., and Bajic, V. B. (2015). DWFS: A wrapper feature selection tool based on a parallel genetic algorithm. PLOS ONE, 10(2), e0117988. https://doi.org/10.1371/journal.pone.0117988

Wang, Y., He, X., Nie, H., Zhou, J., Cao, P., and Ou, C. (2020). Application of artificial intelligence to the diagnosis and therapy of colorectal cancer. American Journal of Cancer Research, 10(11), 3575–3589.

Yang, F., Wan, Y., Shen, X., Wu, Y., Xu, L., Meng, J., Wang, J., Liu, Z., Chen, J., Lu, D., Wen, X., Zheng, S., Niu, T. and Xu, X. (2023). Application of multi-modality MRI-based radiomics in the pre-treatment prediction of RPS6K expression in hepatocellular carcinoma. Molecular Biomedicine, 4(1), 22. https://doi.org/10.1186/s43556-023-00134-z

Yuan, F., Li, Z., Chen, L., Zeng, T., Zhang, Y.-H., Ding, S., Huang, T. and Cai, Y.-D. (2021). Identifying the signatures and rules of circulating extracellular microRNA for distinguishing cancer subtypes. Frontiers in Genetics, 12, 651610. https://doi.org/10.3389/fgene.2021.651610

Zhi, J., Sun, J., Wang, Z., and Ding, W. (2018). Support vector machine classifier for the prediction of the metastasis of colorectal cancer. International Journal of Molecular Medicine, 41(3), 1419–1426.