International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 25(2), November, 2025, pp 47-57
DOI: https://doi.org/10.3329/ijss.v25i2.85734

Model Selection Strategies for Cancer Prediction from Gene Expression Data: A Beginner's Perspective on Machine Learning

Bandhan Sarker and Md. Matiur Rahaman*

Department of Statistics, Faculty of Science,
Gopalganj Science and Technology University, Gopalganj, 8105, Bangladesh

*Correspondence should be addressed to Md. Matiur Rahaman
(Email: matiur.stat@gmail.com)

[Received July 21, 2025; Accepted September 20, 2025]

Abstract

Microarray gene expression data are often classified by cell line or tumor type to assist in the diagnosis and prediction of human cancer. While microarray analysis demonstrates potential, choosing the most suitable machine learning approach is crucial for accurate cancer diagnosis and prediction. In this beginner's guide, we outline how to select an optimal machine learning model for cancer phenotype prediction by comparing various existing methods. This study used three well-known machine learning methods: linear discriminant analysis, support vector machines, and random forest. To assess prediction performance, several performance metrics were considered, including model prediction accuracy (AC), the area under the curve (AUC), F-measure, the receiver operating characteristic (ROC) curve, and the precision-recall curve (PRC). Microarray gene expression data from two cancer types, leukemia and colon cancer were analysed. A cross-validation process with 100 resampling iterations was implemented to compute average performance measures (APM), ensuring the reliability of the results. Findings emphasize the significance of selecting the right machine learning model for accurate predictions of new samples. The methods employed provided satisfactory results, validated by various APM for both leukemia and colon cancer datasets. Notably, the random forest classifier exhibited the best performance in cancer prediction.

Keywords: Machine learning (ML), Microarray gene expression, Support vector machine (SVM), linear discriminant analysis (LDA), Random Forest (RF).

AMS Classification: 62P10, 68T09.

1. Introduction

There is a need to develop analytical methodologies to analyse and to exploit the information contained in gene expression data (Alon et al., 1999). Due to the large number of genes and the complexity of biological networks, statistical tools serve as valuable exploratory techniques for analysing this dataset (Gillani et al., 2014, Namani et al., 2018, Rahaman et al., 2025). Microarray gene expression data consists of a small number of sample with a large number of genes, and these

genes are often highly correlated. Therefore, this extensive dataset presents significant challenges for analysis.

It has been reported in the literature that microarray studies have increased exponentially, and machine learning (ML) methods have become widely accepted for the diagnosis and classification of human cancers. Numerous supervised and unsupervised ML techniques are available. Some supervised ML techniques include linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), naïve Bayes (NB), Gaussian process classification (GPC), support vector machines (SVM), artificial neural networks (ANN), logistic regression (LR), decision trees (DT), AdaBoost (AB), and random forest (RF). Unsupervised learning techniques include k-means clustering, fuzzy neural networks, hierarchical clustering, self-organizing maps (SOM), (FNN), and others (Hanczar et al., 2003, Gillani et al., 2014).

Cancer prediction offers an unbiased and general approach for developing prognostic tests, provided there is a collection of tumor samples with known outcomes. The primary objective of cancer prediction is to identify factors associated with future clinical outcomes, such as drug response or survival. Among the various methods available, an optimized approach is essential for accurate cancer diagnosis. To guide the selection of an appropriate ML model or the development of a new one, it is necessary to evaluate and compare the performance of ML methods for cancer prediction.

In this study, we compared three well-known supervised ML methods: linear discriminant analysis (LDA), support vector machines (SVM), and random forest (RF). LDA is employed in statistics and machine learning to find a linear combination of features for classifying data into two or more classes (Anderson, 2003; Johnson and Wichern, 2007, Rahaman and Mollah, 2019). SVM is a widely used ML method across various fields, including cancer diagnosis, bioinformatics, image classification, feature selection, and text mining. Its popularity arises from a solid mathematical foundation. SVM is effective in handling high-dimensional datasets and performs well in nonlinear classification through the use of kernel functions. (Suykens and Vandewalle, 1999). RF is an ensemble learning method for regression and classification in machine learning, involving the construction of multiple decision trees through bootstrap aggregation (Breiman, 2001). RF operates based on the predictions of these tree structures. During the training process, the model builds multiple individual decision trees, and the prediction from all these trees are combined to arrive at a final decision. One of the key benefits of RF is its effectiveness with categorical data, as it utilizes a majority voting mechanism to establish the outcome by weighing the results from each tree.

Several performance indices were used to select the best machine learning model for accurately classifying new or test samples. Prediction accuracy (AC), receiver operating characteristic (ROC) curves, and the area under the ROC curve (AUC) are widely regarded as key measures of the performance of ML algorithms (Statnikov et al., 2005, Chen et al., 2007, Pirooznia et al., 2008). The precision-recall (PR) curve illustrates the sensitivity of classifiers in cases of moderate to large class imbalance, allowing for an accurate and intuitive interpretation of classifier performance (Saito and Rehmsmeier, 2015). The F-measure maintains a balance between precision and recall for a classifier (Ahamad et al., 2020). Together, accuracy, AUC, F-measure, ROC curves, and PR curves allow us to assess the best classification models for gene expression studies.

This paper, presents the materials and methods of the study in Section 2, while Section 3 provides the analysis results and a detailed discussion. Finally, Section 4 presents the conclusion.

2. Methodological framework and materials used

In this study, we propose a foundational machine learning (ML) model selection framework for beginners interested in cancer prediction based on gene expression studies. Figure 1 illustrates the proposed framework for selecting an ML model. The pseudocode representation is as follows:

Pseudocode representation

```
function ML_Model_Selection(microarray_data):
# Step I: Process the microarray dataset
   processed_data = Preprocess(microarray_data)

# Step II: Identify most significant differentially expressed genes (θ), where θ < n_k; n is the number of samples/patients and k is the sample groups/classes.
   significant_genes (θ) = Identify_DEGs(processed_data)

# Initialize models and performance metrics
   models = [M1, M2, M3, ..., Mα]
   performance_metrics = {}

 # Step III: Sample and split the dataset
   for model in models:
      for i in range(I):  # Repeat for I iterations
         train_data, test_data = Sample_and_Split(processed_data)
         trained_model = Train_Model(model, train_data, significant_genes)

 # Calculate performance measurements (PM)
         pm = Evaluate_Model(trained_model, test_data)
         if model not in performance_metrics:
            performance_metrics[model] = [ ]
         performance_metrics[model].append(pm)

  # Step IV: Calculate average performance measures (APM)
   average_performance = {}
   for model in models:
      average_performance[model] = Average(performance_metrics[model])

  # Step V: Select the optimal model
   optimal_model = Select_Optimal_Model(average_performance)

   return optimal_model
```

Explanation of the Pseudocode

Preprocess: Clean and normalize the microarray data.
Identify_DEGs: Function to find significant differentially expressed genes based on the criteria provided.
Sample_and_Split: Randomly sample and split the dataset into training and testing sets.
Train_Model: Train the specified machine learning model on the training set using the selected significant genes.
Evaluate_Model: Assess the model's performance on the test set and return performance metrics.
Average: Calculate the average performance measures for each model.

Select_Optimal_Model: Choose the model with the best average performance for future predictions.

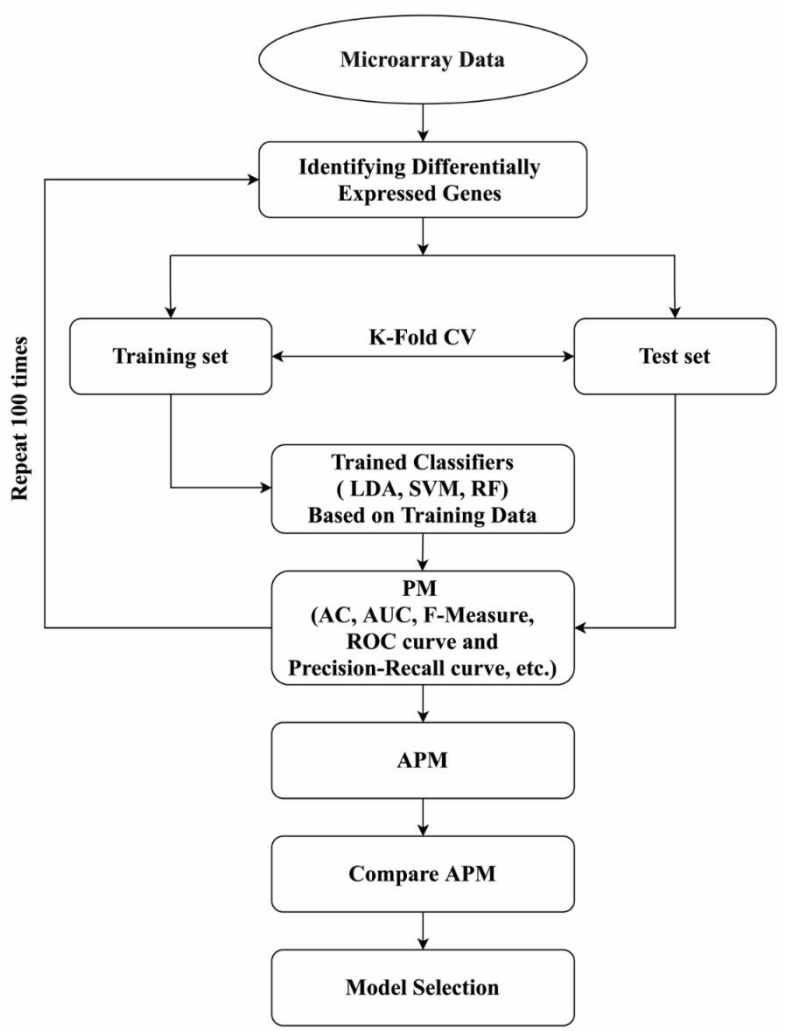

Figure 1: An optimize machine learning model selection framework for cancer predation.

2.1 Datasets description

In this work, we applied machine learning (ML) methods to two microarray gene expression datasets: the Leukemia dataset and the Colon cancer dataset.

2.1.1 Leukemia dataset

Golub et al. (1999) introduced gene expression monitoring as a method to differentiate between two types of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). These leukemia types are classified based on their origins: lymphoid, associated with lymph or lymphatic tissue, and myeloid, linked to bone marrow. ALL can be further categorized into B-cell

and T-cell subtypes. The dataset comprises 7,128 genes and includes 72 samples, with 47 representing ALL and 25 representing AML. This dataset is accessible at http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv.

2.1.2 Colon dataset

We utilized gene expression data from microarray experiments involving colon tissue samples (Alon et al., 1999). This dataset comprises 62 samples, including 40 tumor tissues labeled as 2 and 22 normal tissues labeled as 1, totaling 2,000 genes. The dataset is accessible at http://microarray.princeton.edu/oncology/. Additionally, these datasets are also available in various R packages.

2.2 Differential gene identification

Identifying a set of genes that can act as class predictors is crucial. The t-test, a parametric statistical method, is employed to assess the difference between the means of two samples (cancer and control). This test statistic is used to identify DEGs that might be suitable for class prediction from the datasets in the context of the classification problem under the test hypothesis:

$$H_0 : \mu_{1i} = \mu_{2i} \text{ vs. } H_1 : \mu_{1i} \neq \mu_{2i}$$

and the test statistic is:

$$t = \frac{|\mu_{1i} - \mu_{2i}|}{\sqrt{\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}}} ; i = 1,2,3, \ldots \ldots \ldots G.$$

where, μ_{1i} and μ_{2i} are the means of the two classes, and s_{1i}^2 and s_{2i}^2 are the variances, respectively. n_1 is the sample size of the first group and n_2 is the sample size of the second group, G is the number of genes in the microarray datasets. The test statistic $t \sim t_{n_1+n_2-2}$ follows a *t*-distribution with (n_1+n_2-2) degrees of freedom.

2.3 Cross-validation

Cross-validation (CV) is a resampling method used to assess the performance of ML models on small data samples. The method has one main parameter called k, which indicates the number of groups into which the data sample is divided. When a specific value for k is selected, it is called k-fold cross-validation (for example, k = 10 is 10-fold cross-validation). Braga-Neto and Dougherty explained that cross-validation is mainly applied in ML to estimate how well a machine learning model performs on unseen data (Braga-Neto and Dougherty, 2004). The results of a *k*-fold cross-validation are usually summarized by taking the average of the model's skill scores.

2.4 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) maximizes the separability between classes/groups (say, cancer and control). The mathematical formula for LDA is expressed as:

$$\boldsymbol{X}^T \boldsymbol{\xi}^{-1} (\overline{\boldsymbol{X_1}} - \overline{\boldsymbol{X_2}}) - \frac{1}{2} (\overline{\boldsymbol{X_1}} + \overline{\boldsymbol{X_2}})^T \boldsymbol{\xi}^{-1} (\overline{\boldsymbol{X_1}} - \overline{\boldsymbol{X_2}}) \geq \alpha,$$

where $\boldsymbol{X}$ is a data matrix, $\overline{\boldsymbol{X_1}}$ is the mean vector for the first group, and $\overline{\boldsymbol{X_2}}$ is the mean vector for the second group, $\boldsymbol{\xi}$ is the sample variance-covariance matrix, $\boldsymbol{X^T}$ is the transpose of $\boldsymbol{X}$, and α is the decision boundary threshold, which can be 0, greater than 0, or less than 0. When α = 0, classification results are similar; otherwise, if the value of α > 0, the instance will be classified into

the cancer group and vice versa (Anderson, 2003; Johnson and Wichern, 2007, Rahaman and Mollah, 2019).

2.5 Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning technique used in machine learning (Suykens and Vandewalle, 1999). Let the dataset S consist of observations $x_1, x_2, ..., x_p \in X$ and associated $y_1, y_2, ..., y_n \in y$. A separating hyperplane classifier method is to learn a function $f: X \rightarrow y$ from S that is used to predict the label of any new observation $x \in X$ by $f(x)$ and classified into two classes as $y \in \{-1, +1\}$. Then a separating hyperplane has the property that $W^T X + b > 0$ if $y_i = 1$ and $W^T X + b < 0$ if $y_i = -1$.

2.6 Random Forest (RF)

The Random Forest (RF) algorithm uses an ensemble of classification trees (Breiman, 2001). In RF, a classification tree is constructed by the bootstrap sample of the data, and in each split, the candidate set of variables includes a random subset of the available variables. The RF output represents the mode of the classes predicted by the individual trees. In this study, 500 trees were used.

R packages MASS, e1071, and *randomForest* were utilized for the LDA, SVM, and RF classifiers, respectively.

2.7 Performance measurement

Performance measurements were used to evaluate the models. The performance measures are as follows:

- **True Positive (TP):** An outcome in which the model accurately predicts the positive class.
- **True Negative (TN):** An outcome in which the model accurately predicts the negative class.
- **False Positive (FP):** An outcome in which the model mistakenly predicts the positive class.
- **False Negative (FN):** An outcome in which the model mistakenly predicts the negative class.

Using these measurement values, accuracy, precision, recall, F-measure, and AUC were calculated, and ROC and precision-recall (PR) curves were generated. Therefore,

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F-measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The score of each measure ranges from 0 to 1. A prediction model is considered good if it produces the highest scores across these measures.

3. Results and discussion

In this study, we assessed the performance of three widely used machine learning methods for biological classification and prediction (Hanczar et al., 2003; Pirooznia et al., 2008; Chen et al.,

2013; Gillani et al., 2014; Rahaman et al., 2015; Cui et al., 2019; Rahaman et al., 2019). We used two real datasets to demonstrate our approach to model selection, focusing on classification accuracy and precision. We removed equally expressed genes (EEGs) because they do not aid in class prediction. Using a *t*-statistic with a significance level of $p \leq 0.001$, we identified 726 differentially expressed genes (DEGs) in the Leukemia dataset and 186 DEGs in the Colon dataset. Figure 2A shows the DEGs in the Leukemia dataset, while Figure 2B shows the DEGs in the Colon dataset. Both figures depict the patterns of differential gene expression. In Figure 2, the data are arranged in a grid where each row represents a gene and each column represents a sample, patient, or individual. The color intensity indicates changes in gene expression. Complementary Hierarchical Clustering (CHC) was applied to the DEGs to reveal clustered expression patterns. The gene sets associated with each pattern relate to specific phenotypes. Within each pattern, the gene sets show partial redundancy and similar expression levels; however, they differ across patterns (Hanczar et al., 2003). These differences can be further validated through functional analysis of the genes. To address issues of dimensionality and computational complexity, we selected the top 15 DEGs from each cancer dataset based on their rank according to the *p*-values shown in Figures 3A and 3B for the Leukemia and Colon datasets, respectively, and applied machine learning methods to these top-ranked genes for cancer prediction.

The performance measures, including true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), classification accuracy (AC), area under the ROC curve (AUC), and F-measure, were calculated for each method. Using TP, FP, TN, and FN, ROC and precision-recall (PR) curves were created. We employed a 10-fold cross-validation (CV) technique and repeated this process 100 times to calculate average performance metrics (APM) for each machine learning classification model.

Figure 4 shows the average prediction accuracy from the 10-fold CV with 100 repetitions for LDA, SVM, and RF. The prediction accuracy for each method ranged from 85% to 96% per iteration, with RF consistently outperforming LDA and SVM in the Leukemia dataset (Figure 4A). Similarly, Figure 4B shows that the accuracy ranged from 83% to 95% per iteration for the Colon dataset, with RF generally achieving higher accuracy than LDA and SVM. The overall average accuracy for each method is summarized in Table 1.

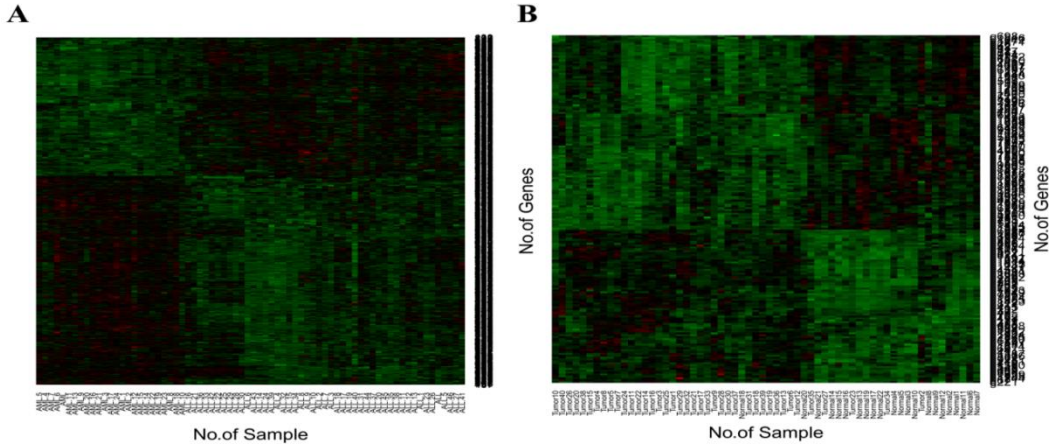

Figure 2: Differentially expressed genes. (A) Leukemia data, and (B) Colon data.

The prediction accuracies were also evaluated using the receiver operating characteristic (ROC) curve (Figures 5A and 5B) through cross-validation. All machine learning methods showed high accuracy, with areas under the curve (AUCs) of 0.93 or higher for the Leukemia dataset and 0.91 or higher for the Colon dataset. Among the ML methods, the RF-based classifier achieved the highest accuracy, with AUCs of 0.96 and 0.93 for the ROC curves, compared to LDA and SVM (Table 1). From Table 1, we observed that the F-measures for each ML method were also high—greater than 0.92 for Leukemia and 0.90 for Colon—with LDA and SVM yielding lower F-measures than the RF-based classifier (F-measures of 0.95 and 0.94, respectively). The precision-recall (PR) curve is useful for evaluating binary classifiers (Saito and Rehmsmeier, 2015). Figures 5C and 5D illustrate the PR curves for each method, showing the performance of the models. We found that all models performed well, with RF showing the best results.

The results of this study indicate that LDA had a prediction accuracy (AC) of over 91% for the Leukemia dataset and 87% for the Colon dataset on average. SVM achieved an average prediction accuracy of 92% for the Leukemia dataset and 87% for the Colon dataset. In contrast, the RF-based classifier demonstrated AC of 93% and 89% for the Leukemia and Colon datasets, respectively. Overall, we observed that all machine learning methods had high and satisfactory accuracy.

From the ROC and PR curves, along with AUC and F-measure, we conclude that the ML models were well-fitted and accurately classified tumors and patients. The performance of the RF-based classifier was superior to that of LDA and SVM, while SVM outperformed LDA. However, the performance measures (PM) for LDA and SVM, as well as SVM and RF, were quite close to each other, except RF and LDA. Notably, there was no significant difference between the performances of SVM and RF. Based on various performance metrics, we conclude that the RF-based classifier is the most effective among the three ML methods. This RF classifier can be used for classifying new samples if the datasets come from the same platform.

The limitations of our study include the understanding that classification performance depends on significant features identified in the literature. Therefore, our analysis would benefit from a more robust feature selection method (Mollah et al., 2015). Additionally, we included only three ML methods in this study; exploring more ML methods during model selection would be valuable. The best ML model identified through the provided approach will reduce the likelihood of misclassification of new samples and improve the accuracy of cancer classification.

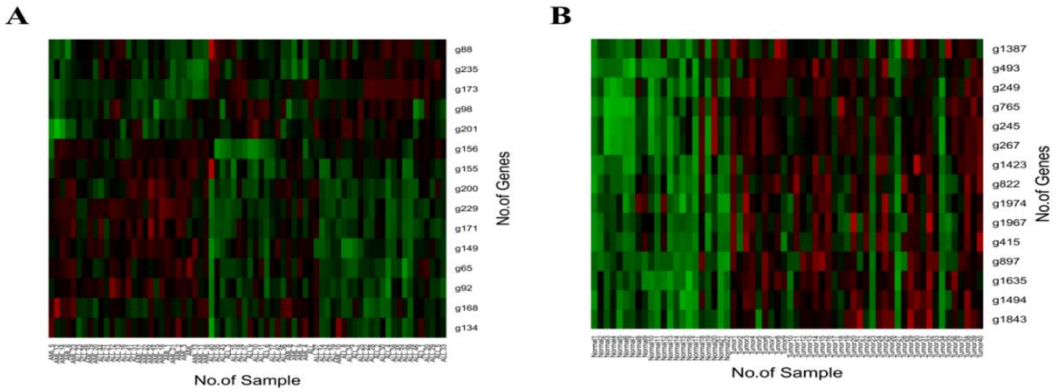

Figure 3: Top 15 differentially expressed genes. (A) Leukemia data, and (B) Colon data.

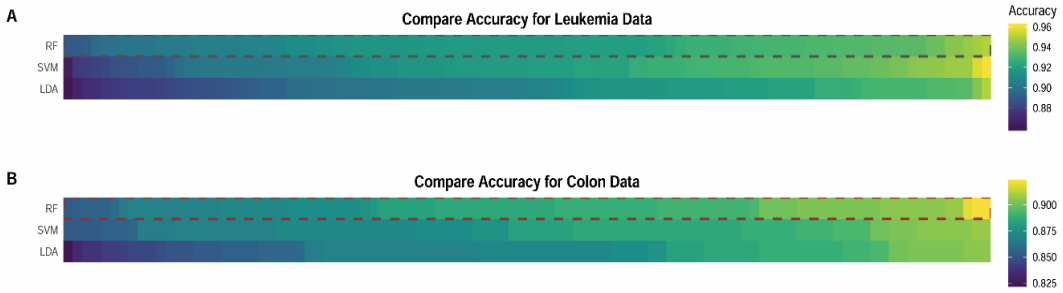

Figure 4: Average prediction accuracy of LDA, SVM, and RF obtained by 10-fold CV with 100 repetitions. (A) Leukemia data, and (B) Colon data.

Table 1: Accuracy (AC), Area under the ROC curve (AUC), and F-measure.

	Leukemia data			Colon data		
Methods	**AC**	**AUC**	**F-measure**	**AC**	**AUC**	**F-measure**
LDA	0.91	0.93	0.92	0.87	0.91	0.90
SVM	0.92	0.95	0.94	0.87	0.93	0.92
RF	**0.93**	**0.96**	**0.95**	**0.89**	**0.93**	**0.94**

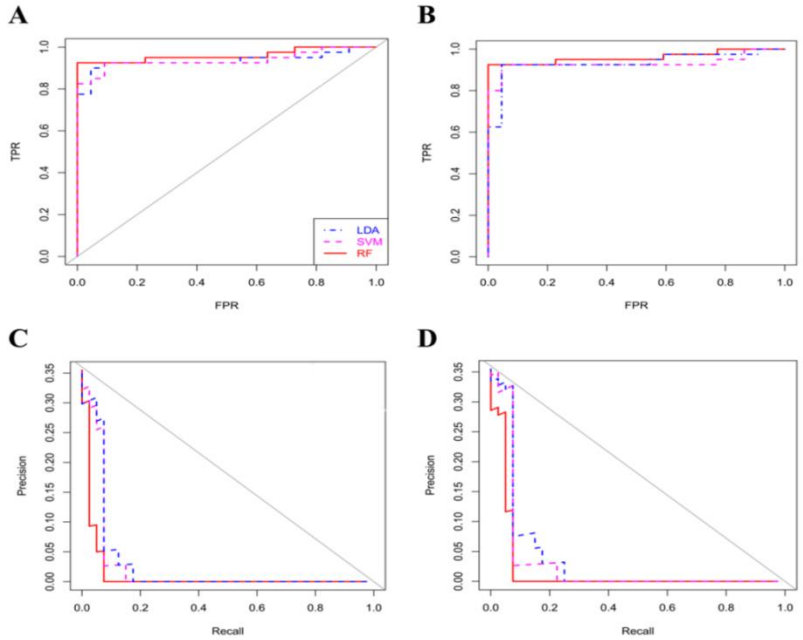

Figure 5: ROC and PR curve. (A) ROC curve of Leukemia data, (B) ROC curve of Colon data, (C) PR curve of Leukemia data, and (D) PR curve of Colon data.

Before making accurate cancer predictions, we want to emphasize some important points:

Identification of Significant Genes/Biomarkers: There is a need to identify more significant genes or biomarkers.

Evaluation of ML Models: To guide the selection of an appropriate ML model or the development of a new one, it is essential to evaluate and compare the performance of ML methods based on training datasets and their respective groups or classes.

Comprehensive Performance Metrics: Most studies on cancer prediction or classification in the literature have only assessed model performance using accuracy, which is not sufficient. Therefore, to select the best ML model, it is necessary to apply more comprehensive metrics such as recall, precision, F-measure, ROC curve, AUC, and others.

Application of Optimized Model: An optimized machine learning model can then be used to classify test samples or datasets with unseen groups or classes.

4. Conclusion

In summary, we provided a computational framework for selecting machine learning models based on previous research, targeting beginners interested in cancer prediction and classification. This optimized machine learning approach will be executed accurately and produce results aligned with established principles. In this study, we used three supervised learning algorithms: Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and Random Forest (RF), and found their performances to be quite strong. Further investigations could explore additional algorithms in model selection to enhance classification accuracy for new samples, potentially utilizing advanced statistical methods for gene selection in cancer prediction. Overall, SVM and LDA showed generally good prediction accuracy; however, the RF classifier demonstrated superior performance for both Leukemia and Colon cancer datasets. This indicates that among the considered machine learning candidates, RF is the most appropriate choice for the final classifier when classifying new samples from datasets referencing the same platform.

Acknowledgments: We would like to thank those who worked on machine learning-based gene expression studies. We also acknowledge AI-powered chatbots and tools for improving the grammar and clarity of this manuscript. The author thanks the reviewer for their valuable insights and for helping to strengthen the arguments presented.

Authors' contributions: BS and MMR developed the idea or concept. The final validation of the paper was done by BS and MMR. The entire process was monitored by MMR.

Conflict of interest: The authors declare no conflict of interest.

References

[1] Ahamad, M., Aktar, S., Uddin, S., Lió, P., Xu, H., Summers, M. A., Quinn, J. M. and Moni, M. A. (2020). A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Systems with Applications. 2020; 113661.

[2] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America. 1999; 96(12):6745-6750.

[3] Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. Wiley Interscience.

[4] Braga-Neto, U. M. and Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? Bioinformatics. 2004; 20(3):374-380.

[5] Breiman, L. (2001). Random forests. Machine Learning. 2001; 45(1):5-32.

[6] Chen, T., Cao, Y., Zhang, Y., Liu, J., Bao, Y., Wang, C., Jia, W. and Zhao, A. (2013). Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evidence-Based Complementary and Alternative Medicine. 2013; 298183.

[7] Chen, X., Zhao, Y., Zhang, Y.-Q. and Harrison, R. (2007). Combining SVM classifiers using genetic fuzzy systems based on AUC for gene expression data analysis. In: International Symposium on Bioinformatics Research and Applications. 2007. Springer, pp. 496-505.

[8] Cui, S., Wu, Q., West, J. and Bai, J. (2019). Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease. PLoS Computational Biology. 2019; 15(8):e1007264.

[9] Gillani, Z., Akash, M. S., Rahaman, M. D. and Chen, M. (2014). CompareSVM: supervised, support vector machine (SVM) inference of gene regularity networks. BMC Bioinformatics. 2014; 15:395.

[10] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R. and Caligiuri, M. A. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439):531-537.

[11] Hanczar, B., Courtine, M., Benis, A., Hennegar, C., Clément, K. and Zucker, J.-D. (2003). Improving classification of microarray data using prototype-based feature selection. ACM SIGKDD Explorations Newsletter. 2003; 5(2):23-30.

[12] Johnson, R. A., and Wichern, D. W. (2007). Applied multivariate statistical analysis. Sixth edition, Prentice-Hall.

[13] Mollah, M. M. H., Jamal, R., Mokhtar, N. M., Harun, R. and Mollah, M. N. H. (2015). A hybrid one-way ANOVA approach for the robust and efficient estimation of differential gene expression with multiple patterns. PLoS One. 2015; 10(9):e0138810.

[14] Namani, A., Rahaman, M. M., Chen, M. and Tang, X. (2018). Gene-expression signature regulated by the KEAP1-NRF2-CUL3 axis is associated with a poor prognosis in head and neck squamous cell cancer. BMC Cancer. 2018; 18(1):46.

[15] Pirooznia, M., Yang, J. Y., Yang, M. Q. and Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008; 9 Suppl 1:S13.

[16] Rahaman, M. M., Sarker, B., Alamin, M. H., and Ferdousi, F. (2025). An integrated approach for key gene selection and cancer phenotype classification: Improving diagnosis and prediction. Computers in Biology and Medicine. 2025; 196:110687.

[17] Rahaman, M. M., Ahsan, M. A. and Chen, M. (2019). Data-mining techniques for image-based plant phenotypic traits identification and classification. Scientific Reports. 2019; 9(1):19526.

[18] Rahaman, M. M., Chen, D., Gillani, Z., Klukas, C. and Chen, M. (2015). Advanced phenotyping and phenotype data analysis for the study of plant growth and development. Frontiers in Plant Science. 2015;6: 619.

[19] Rahaman, M. M. and Mollah, M. N. H. (2019). Robustification of Gaussian Bayes Classifier by the Minimum β-Divergence Method. Journal of Classification, 36, 113–139.

[20] Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015; 10(3):e0118432.

[21] Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D. and Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics. 2005; 21(5):631-643.

[22] Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural Processing Letters. 1999; 9(3):293-300.