Global Medical Genetics 12 (2025) 100079

Contents lists available at ScienceDirect

Global Medical Genetics

journal homepage: www.keaipublishing.com/en/journals/gmg/

Machine learning approach to identify significant genes and classify cancer types from RNA-seq data

Sultana Akter[a], Ridwan Olamilekan Adesola[b,*], Shreya Basnet[c]

[a] College of Medicine and Life Sciences, Biomedical Sciences Concentrate Bioinformatics, University of Toledo, Ohio, USA
[b] Department of Veterinary Medicine and Surgery, University of Missouri, Columbia, MO, USA
[c] Department of Electric Engineering and Computer Science, University of Missouri, Columbia, MO, USA

ARTICLE INFO

Keywords:
Cancer
Diagnosis
Machine learning
RNA seq

ABSTRACT

Cancer remains a leading cause of morbidity and mortality worldwide, with nearly 10 million deaths reported in 2022. In the United States, more than 618,000 deaths are projected to occur in 2025. Traditional methods for identifying cancer types are often time-consuming, labor-intensive, and resource-demanding, highlighting the need for efficient alternatives. This study aimed to evaluate machine learning algorithms on RNA-seq gene expression data to identify statistically significant genes and classify cancer types. We retrieved the PANCAN RNA-seq dataset from the UCI Machine Learning Repository and assessed eight classifiers—Support Vector Machines, K-Nearest Neighbors, AdaBoost, Random Forest, Decision Tree, Quadratic Discriminant Analysis, Naïve Bayes, and Artificial Neural Networks. Model performance was validated using a 70/30 train-test split and 5-fold cross-validation. Among the tested models, the Support Vector Machine achieved the highest classification accuracy of 99.87 % under 5-fold cross-validation. These findings demonstrate the potential of machine learning to efficiently analyze RNA-seq data, facilitate biomarker discovery, and support the development of personalized cancer diagnostics and treatment strategies.

Introduction

Cancer comprises a heterogeneous group of diseases characterized by uncontrolled cell growth in different tissues and organs. It is the second leading cause of death worldwide, accounting for over 9.7 million deaths annually [1,2]. In the United States, more than two million new cases are diagnosed each year [1]. Accurate identification of cancer type is critical, as it directly influences treatment decisions and patient survival outcomes [3]. Early detection through clinical screening remains valuable; however, conventional methods for cancer diagnosis and classification are often labor-intensive, time-consuming, and costly [4].

Aberrant gene expression is a hallmark of cancer [4]. RNA sequencing (RNA-seq) enables comprehensive profiling of gene expression, providing critical insights for cancer diagnosis and molecular characterization. With the advancement of high-throughput RNA-seq technologies, large-scale datasets such as the RNA-Seq (HiSeq) PANCAN resource have become available to the research and medical communities. Leveraging these data for precise cancer type classification, however, requires robust computational methods.

* Corresponding author.
E-mail addresses: sakter2@rockets.utoledo.edu (S. Akter), ra933@umsystem.edu (R.O. Adesola), sb89c@umsystem.edu (S. Basnet).

https://doi.org/10.1016/j.gmg.2025.100079
Received 23 August 2025; Received in revised form 29 September 2025; Accepted 1 October 2025
Available online 10 October 2025

Table 1
Data Overview.

Variable Name	Description	Levels
Class	Cancer Type (BRCA, KIRC, LUAD, PRAD, COAD)	1 = BRCA, 2 = KIRC, 3 = LUAD, 4 = PRAD, 5 = COAD
Features	Gene expression value (total of 20,531 genes)	Higher value indicates more active gene expression, and lower indicates less activity

Machine learning (ML) has gained considerable attention in recent years for its ability to improve disease diagnosis by enhancing accuracy, efficiency, and reproducibility [5,6]. While several studies have applied ML approaches to classify cancers—including breast, lung, prostate, and brain cancers—many of these efforts have relied on clinical imaging or microarray data [7]. Relatively few studies have harnessed RNA-seq data, which presents unique challenges due to its high dimensionality, gene–gene correlations, and potential noise, often combined with limited sample sizes. These challenges can lead to overfitting and multicollinearity in predictive models.

To address these limitations, the present study applies feature selection strategies, including Lasso regression and Random Forest, in combination with machine learning classifiers to identify statistically significant genes and accurately classify cancer types from RNA-seq data. This integrative framework aims to improve biomarker discovery and support the development of more efficient cancer diagnostic strategies.

Materials and methods

Data

The dataset used in this study originates from the UCI Machine Learning Repository (UCI Machine Learning Repository's "Gene Expression Cancer RNA-Seq" dataset [8]) and is titled the "Gene Expression Cancer RNA-Seq" dataset. It is based on RNA sequencing data provided by The Cancer Genome Atlas (TCGA), one of the most comprehensive cancer genomics databases. RNA-Seq was conducted using the Illumina HiSeq platform, which provides high-throughput, accurate quantification of transcript expression levels.

The dataset consists of 801 cancer tissue samples with number of 20,531 genes representing five distinct cancer types (BRCA – Breast Cancer, KIRC – Kidney Renal Clear Cell Carcinoma, COAD – Colon Adenocarcinoma, LUAD – Lung Adenocarcinoma, LUAD – Lung Adenocarcinoma) (Table 1; Fig. 1). Each sample includes expression data for thousands of genes, enabling large-scale analysis of transcriptional activity across different tumor types. One notable characteristic of this dataset is class imbalance, meaning that the number of samples varies across cancer types, which may introduce bias in predictive modeling. Therefore, preprocessing techniques such as down-sampling or data balancing are often applied before model training. This dataset provides a rich foundation for

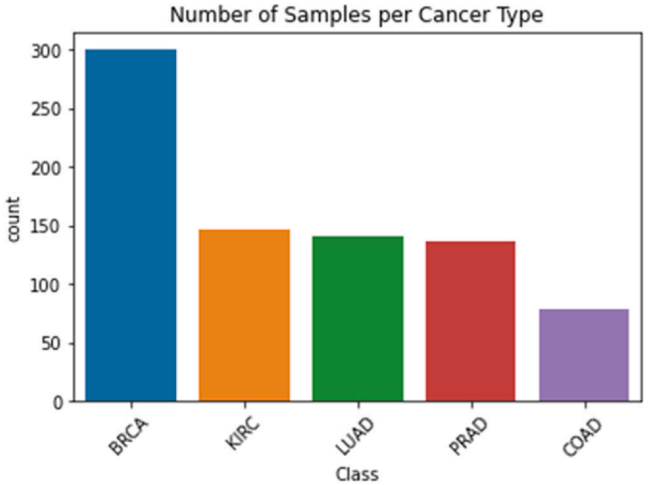


Fig. 1. Frequency of cancer types.

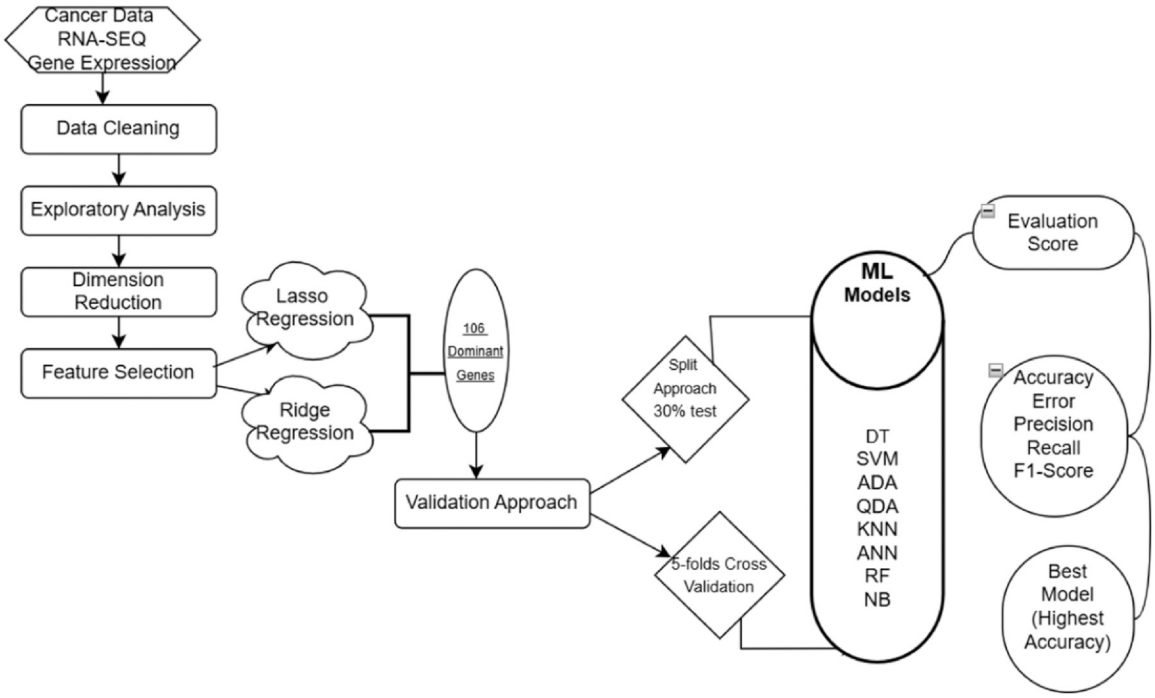


Fig. 2. Analysis pipeline.

investigating cancer-related gene expression patterns, performing biomarker discovery, and developing classification models for tumor identification.

In addition, we utilized data from Brain Cancer Gene Expression (CuMiDa: https://www.kaggle.com/datasets/brunogrisci/brain-cancer-gene-expression-cumida) for external validation of our models.

Features downsampling

RNA-seq gene expression data typically involves a large number of genes relative to a small sample size, with high correlation and significant noise. Various supervised and unsupervised machine learning (ML) techniques have been employed to identify the most relevant genes, though challenges remain due to data complexity [3]. So, in our analysis we utilized Lasso and Ridge Regression algorithms to find the best features (dominant genes) from our data (Fig. 2).

Feature selection statistical methods

Ridge Regression: It is a linear modeling technique commonly used in RNA-seq data analysis to address multicollinearity among genetic markers and identify dominant genes amid a high level of noise. By applying L2 regularization (shrinkage penalty), it penalizes large coefficients to reduce the risk of overfitting, making it well-suited for high-dimensional genomic datasets. This approach effectively balances bias and variance, offering stable and reliable predictions such as breeding values while efficiently handling thousands of genetic features, which makes it a valuable tool in machine learning applications for genomic analysis.

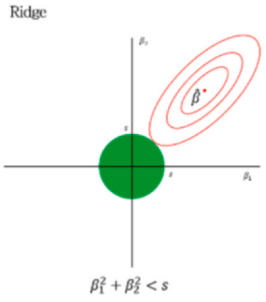
In a standard linear regression model, the goal is to minimize the sum of squared errors (SSE):

$$\sum (y_i - \hat{y}_i)^2$$

where $y_i$are actual values and $\hat{y}_i$ are predicted values.

In Ridge Regression (L2 Regularization), a penalty term is added to control the size of the coefficients:

$$\sum (y_i - \hat{y}_i)^2 + \lambda \Sigma \beta_j^2$$

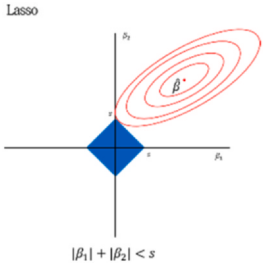


where:

- λ is the regularization parameter (controls the strength of the penalty).
- β_j are the regression coefficients.
- $\Sigma\beta_j^2$ ensures that large coefficients are penalized.

Lasso (Least Absolute Shrinkage and Selection Operator): It is an embedded method that performs feature selection during the model training process. It incorporates regularization by penalizing the absolute magnitude of the regression coefficients. This penalty not only shrinks the coefficients like Ridge but also drives some of them to exactly zero, effectively selecting a subset of relevant features while eliminating the less important ones. Thus, Lasso serves both as a regularization technique and a tool for automatic variable selection within the model.

$$\sum (y_i - \hat{y}_i)^2 + \lambda\Sigma \mid \beta_j \mid$$



In this formulation, the L1 penalty term $\lambda\Sigma \mid \beta_j \mid$ constrains the absolute magnitude of the model coefficients. As a result, many coefficients are shrunk exactly to zero, effectively performing automatic feature selection. This makes Lasso particularly useful when dealing with high-dimensional data where only a subset of features may be informative.

$$\sum (y_i - \hat{y}_i)^2 + \lambda\Sigma \mid \beta_j \mid$$

- This is the Lasso regression cost function combining least squares error with an L1 penalty $\lambda\Sigma \mid \beta_j \mid$.
- The L1 penalty encourages sparsity by shrinking some coefficients exactly to zero.
- Coefficients reduced to zero effectively remove irrelevant features, making Lasso a built-in feature selection method.

Analysis process

Data preprocessing

All the missing values and outliers were checked in the data. The data contain no missing values. Python programming software was utilized for this analysis. All the code used in this study is publicly available at: https://github.com/tonikazic/s25_cptnl_gen/blob/master/shreya_et_al/Shreya%2CSultana%20and%20Ridwan/Project_python_script_breast_cancer.py

Machine learning models

The performance of eight machine learning classifiers was evaluated: Decision Tree, Support Vector Machine, Adaboost, Artificial Neural Network (ANN), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Random Forest, and Naïve Bayes. Briefly, by definition, decision tree (DT) is a non-parametric supervised learning technique for regression and classification. The objective is to build a model that, by utilizing basic decision rules deduced from the data features, predicts the value of a target variable. For classification, they begin at the tree's root and proceed through binary splits based on possible outcomes until they reach

a leaf node and give the final binary result. Support Vector Machine (SVM) is powerful, easy to explain, and versatile, which distinguishes classes with a decision boundary. For this study, we examined two important SVM parameters, cost(C) and gamma. The parameters were set at cost = 1, and gamma = scale. AdaBoost is an ensemble learning algorithm that combines multiple weak classifiers to form a strong predictive model. Artificial Neural Networks (ANNs) are computational models inspired by the structure and functioning of the human brain. ANNs consist of interconnected layers of nodes (also called neurons or units) where each connection has an associated weight. Quadratic Discriminant Analysis (QDA) is a Linear Discriminant Analysis (LDA) variant that separates classes non-linearly. QDA believes each class has its own covariance matrix, unlike LDA. This makes QDA more versatile for datasets without a shared covariance matrix. K-nearest neighbors (KNN) is a method of classifier which is a non-parametric method. The random forest (RF) algorithm is a modified version of the bagging method that combines bagging and feature randomness to generate a collection of decision trees that are not connected with each other. Naïve Bayes (NB) is a family of simple, yet effective probabilistic classifiers based on Bayes' Theorem with a strong (naïve) assumption of conditional independence between features given the class label.

Validation approach

Two validation approaches were used to test the models. They are split- and cross-validation. For the split-validation, the train dataset contains 70 % observations, and the test dataset contains 30 % observations. For cross validation, we utilized *k*-fold cross-validation process. All the classification models were evaluated with 5 5-fold cross-validation.

Statistical evaluation score

The fitted models were evaluated on accuracy score, error rate, precision, recall, and f1 score. Here, we used accuracy scores for each classification model and figured out the highest accuracy score (which means of correct predictions). The confusion matrix gave an accurate score (correct predictions) to calculate the diagonal elements of this matrix. Confusion matrices are tables used to describe the performance of a classification model on a set of test data that is already well-known. There are 4 variables that make up the confusion matrix. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the four possible outcomes. In order to perform an accurate evaluation of the machine learning classifiers, important measures were collected from the confusion matrix. In addition to the correct classification rate or accuracy, other metrics such as True Positive Rate (TPR), False Positive Rate (FPR).

Results and discussions

Data exploratory analysis

Fig. 3A shows principal component analysis (PCA) of the genes expressed. This serves as our initial step, transforming the original high-dimensional data into orthogonal principal components to retain as much variability as possible in a reduced-dimensional space. However, this linear transformation has limitations. In our case, the two-dimensional PCA projection fails to capture correlations in gene expression data, making it difficult to assess whether gene interactions are linear, nonlinear, or more complex. To address this, we employ Kernel PCA, a more advanced technique that utilizes kernel functions to uncover nonlinear relationships that standard PCA might overlook. By mapping the data into a higher-dimensional feature space and then applying PCA, Kernel PCA can detect intricate structures and hidden patterns. Unlike linear PCA, Kernel PCA reveals significant nonlinear patterns in our dataset, indicating that gene expression interactions may involve complex, nonlinear dynamics, an insight with important biological implications. The boxplots in Fig. 3B show top six dominant genes and their distinct expression patterns across tumor types. In particular, some genes (such as gene_4178 and gene_15994) exhibit notably higher expression in BRCA compared to other tumor types. This suggests that these genes may play significant roles in BRCA-related biological processes, such as cell proliferation, tumor progression, or specific molecular signaling pathways. Their differential expression makes them potential biomarkers or therapeutic targets for breast cancer. Fig. 3C illustrates the correlation matrix of sampled gene expressions, where each square represents the correlation coefficient between a pair of genes positioned along the x and y axes. The color intensity of each square denotes the strength and direction of the correlation where red indicates positive correlation, while blue reflects negative correlation. Clusters of red squares suggest groups of genes that are positively correlated and may be co-regulated or functionally related within the same biological pathway. In contrast, blue clusters may point to inversely regulated genes or those involved in distinct biological processes. Examining the intensity patterns can help identify genes with strong inter-relationships, offering valuable insights into potential gene interactions and their functional roles. Additionally, outliers or unique patterns in the heatmap may indicate a typical gene behavior, warranting further investigation. Additionally, outliers or unique patterns in the heatmap may indicate a typical gene behavior, warranting further investigation.

Feature selection

For the final analysis, we initially selected the top 250 genes from both Lasso and Ridge regression analyses (Fig. 4A-B). From these, we identified 106 common genes shared by both the Lasso and Random Forest models, which were used as the final feature set (Fig. 4 C). After downsampling, the resulting dataset comprised 801 samples, 106 genes, and 5 distinct cancer types.

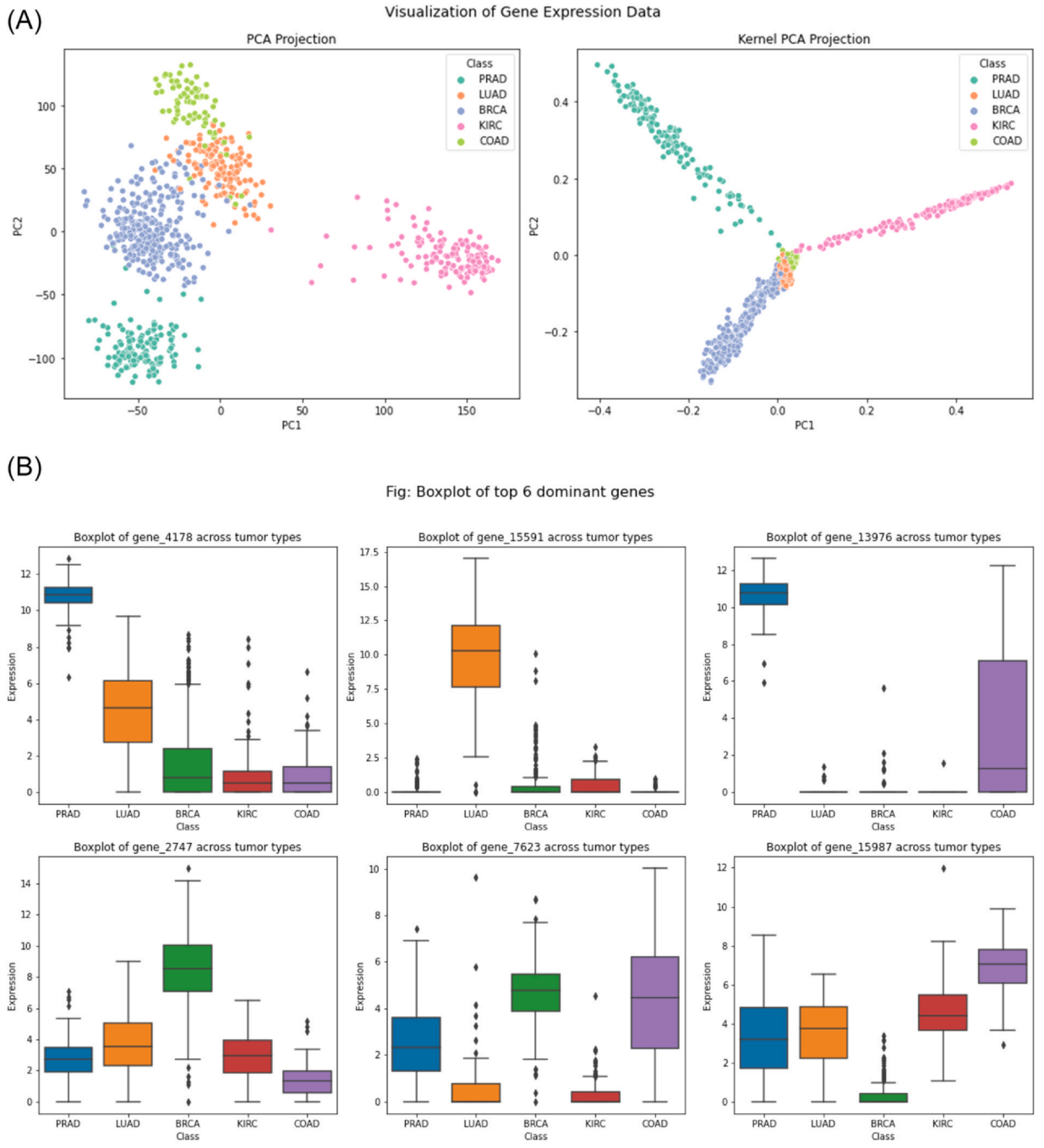


Fig. 3. Exploratory analysis (A) Principal component analysis of genes expressed in the cancer types. (B) Boxplot of top 6 dominant genes. (C) Correlation matrix of genes with their expression values.

Gene expression and interaction

Fig. 5 displays a heatmap of the top 50 genes across tumor types, where yellow indicates high expression and blue represents low expression. Hierarchical clustering is shown through dendrograms along the top and left, grouping genes and tumor samples by expression similarity to shorter branches indicate stronger resemblance. Distinct color blocks reveal co-expressed genes and condition-specific patterns, such as high-expression yellow stripes or broadly expressed green regions. Intense yellow spots amid dark backgrounds may indicate outlier genes with unique expression profiles. This visualization highlights regulatory patterns and potential biomarkers for further biological investigation.

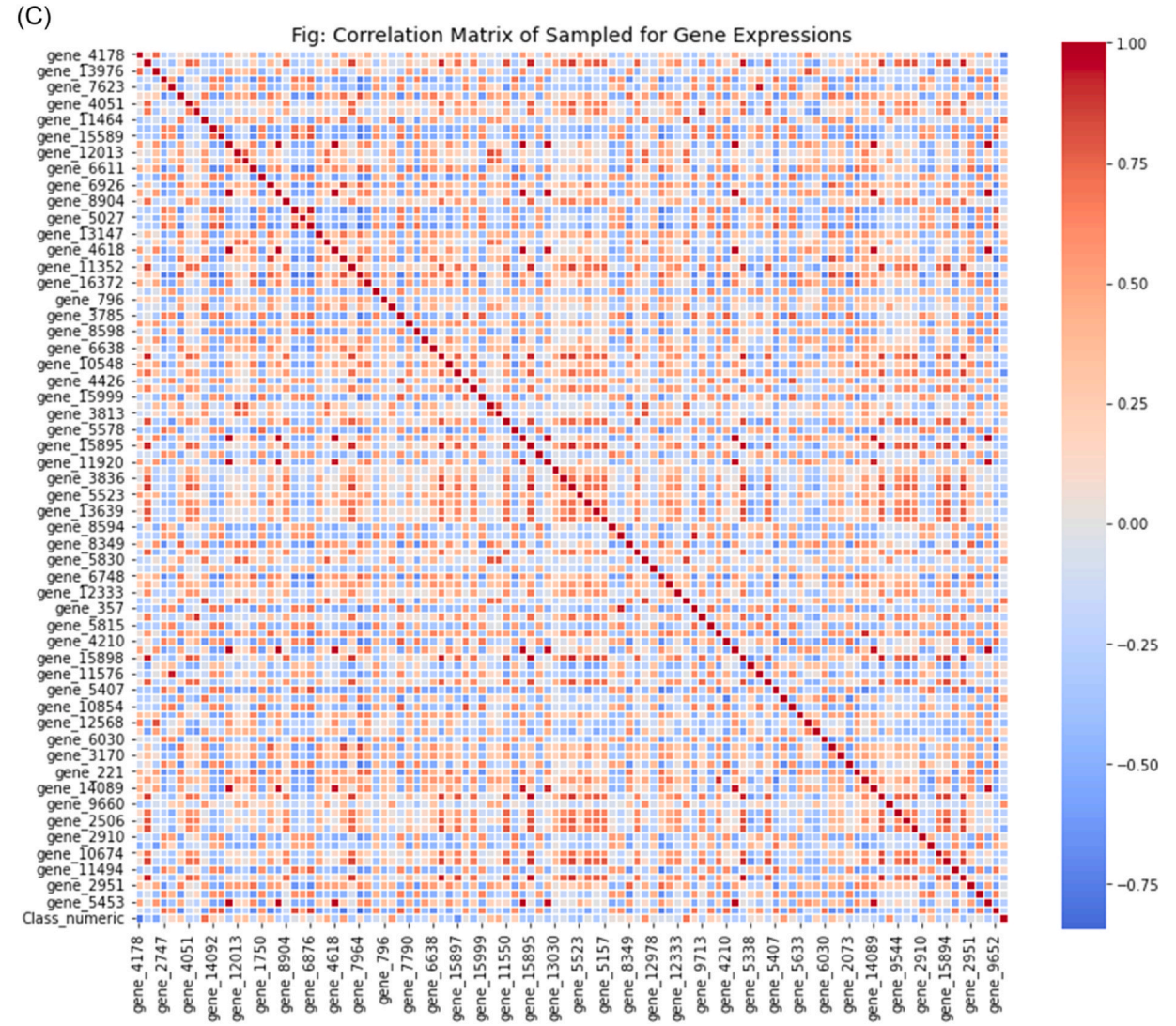


Fig. 3. (*continued*)

The original modularity score of 0.5451 reflects a strong and well-defined community structure within the gene network, indicating that genes within the same module are highly interconnected (Fig. 6A). In contrast, the average modularity of randomized networks is significantly lower at 0.1603, suggesting that the observed structure is not a result of random chance. This is further supported by a highly significant p-value of 0.00001, confirming the robustness of the detected communities. The gene co-expression network was constructed using a correlation threshold of > 0.65, meaning only strongly correlated gene pairs were connected, thereby enhancing the reliability of the network structure (B). The visual and statistical analyses together tell a compelling story: the gene interaction network has important structures that are linked to the biological processes that cause cancer. These findings pave the way for targeted biological pathway analyses and experimental validation to elucidate the roles of these gene communities in cancer biology.

Performance of machine learning models

SVM, ANN, and RF models obtained the highest accuracy scores of 0.9958 each (Table 2). They demonstrated extremely dependable prediction abilities with equivalent precision scores of 0.9955 for SVM and 0.9978 for both ANN and RF. These scores imply that these models minimize false positives and false negatives while producing accurate classifications that are consistent. Conversely, other models with much lower accuracy scores included Decision Tree (DT), AdaBoost (ADA), Quadratic Discriminant Analysis (QDA), and Naïve Bayes (NB). In processing high-dimensional gene expression data, this performance gap demonstrates the superiority of SVM, ANN, and RF and validates their appropriateness for accurate cancer subtype classification.

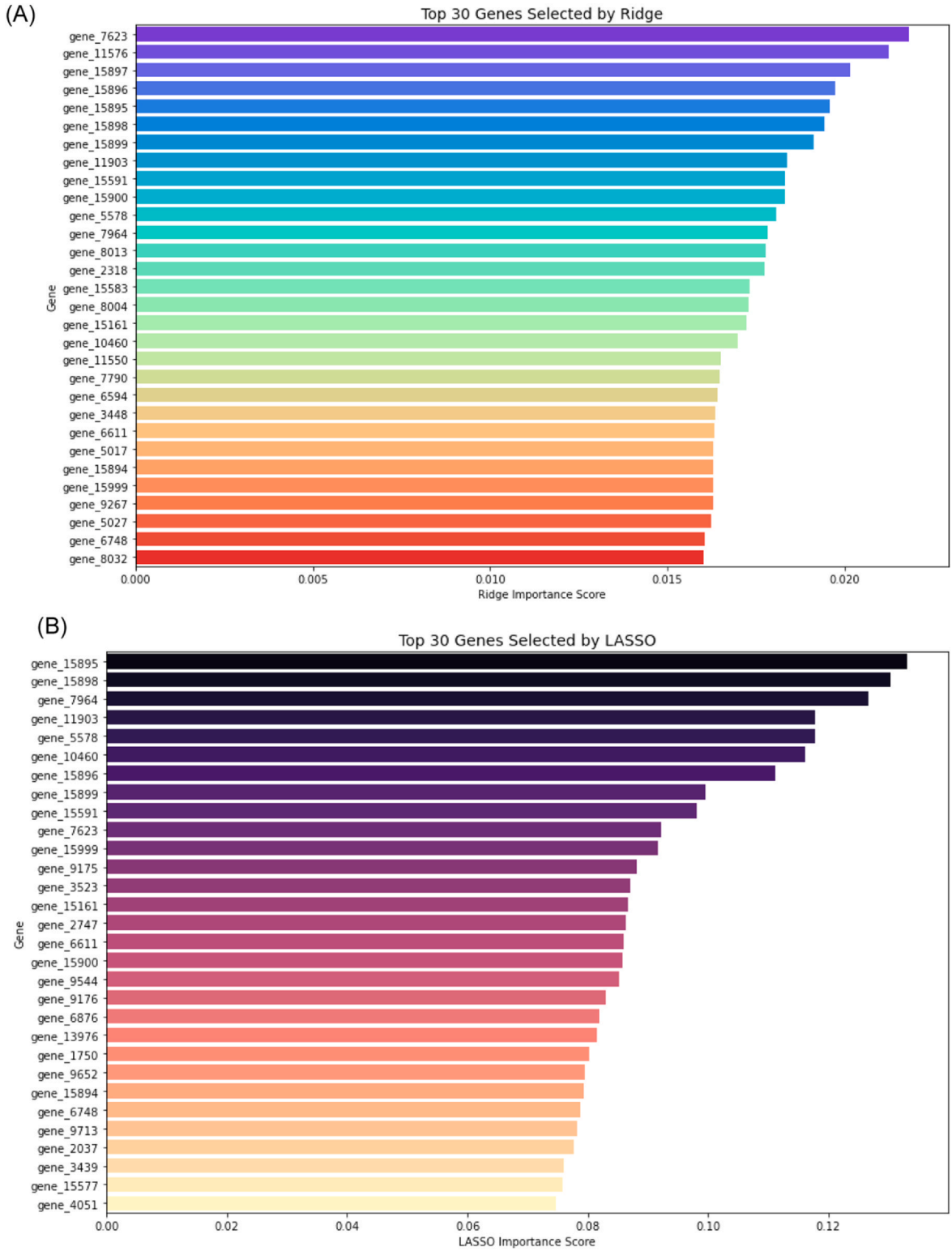


Fig. 4. (A) Most Dominant 30 genes of Ridge. (B) Most Dominant 30 genes of Lasso. (C) Common genes between Lasso and Ridge.

The confusion matrices highlight the stark contrast in performance between the models with the highest and lowest accuracy scores in classifying five cancer subtypes: PRAD, LUAD, BRCA, KIRC, and COAD (Fig. 7). Fig. 7A, The SVM model demonstrates excellent classification capability, achieving perfect accuracy for PRAD, BRCA, KIRC, and COAD, with only a single misclassification

(C)

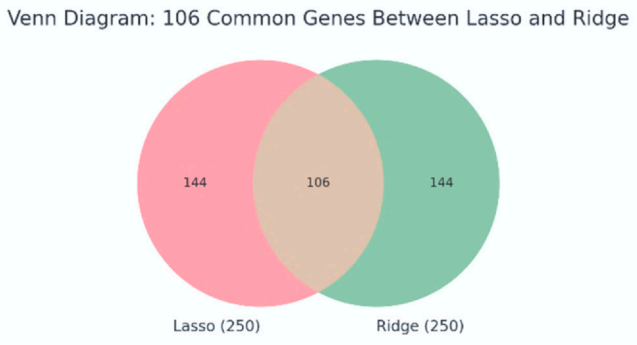


Fig. 4. (*continued*)

where one LUAD sample is incorrectly predicted as KIRC. In contrast, Fig. 7B, the QDA classifier shows substantial misclassifications, particularly mislabeling many PRAD, LUAD, KIRC, and COAD samples as BRCA, despite correctly classifying all BRCA samples. This pattern suggests that QDA struggles to handle the complex boundaries in the dataset, likely due to its assumption of class-specific covariance and inability to model non-linear separations effectively. Overall, SVM significantly outperforms QDA in both accuracy and reliability, making it a more suitable model for high-dimensional RNA-seq data in cancer classification tasks.

A 5-fold cross-validation approach performed a more thorough evaluation by dividing the dataset into five subgroups, four for training and one for testing in each iteration, limiting the possibility of overfitting and enhancing consistency. The SVM model was the most successful, outperforming all other models with an exceptional accuracy score of 0.9987 (Table 3). In comparison, the QDA model performed poorly with an accuracy of 0.4981, showing it is not suitable for this high-dimensional data. Models like KNN, ANN,

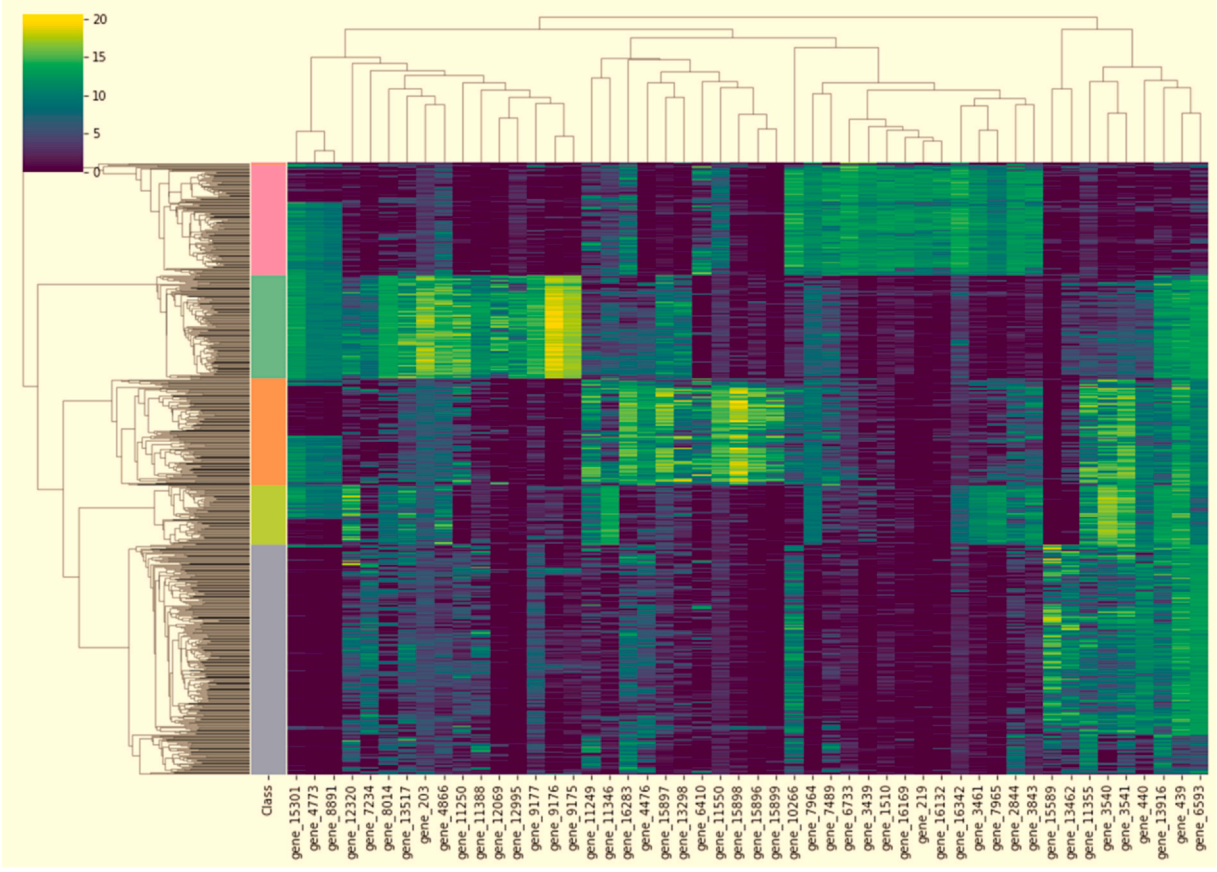


Fig. 5. Heatmap of gene expressions.

(A)

Metric	Value
Original Modularity	0.5451
Avg Randomized Modularity	0.1603
p-value	0.0000

(B)

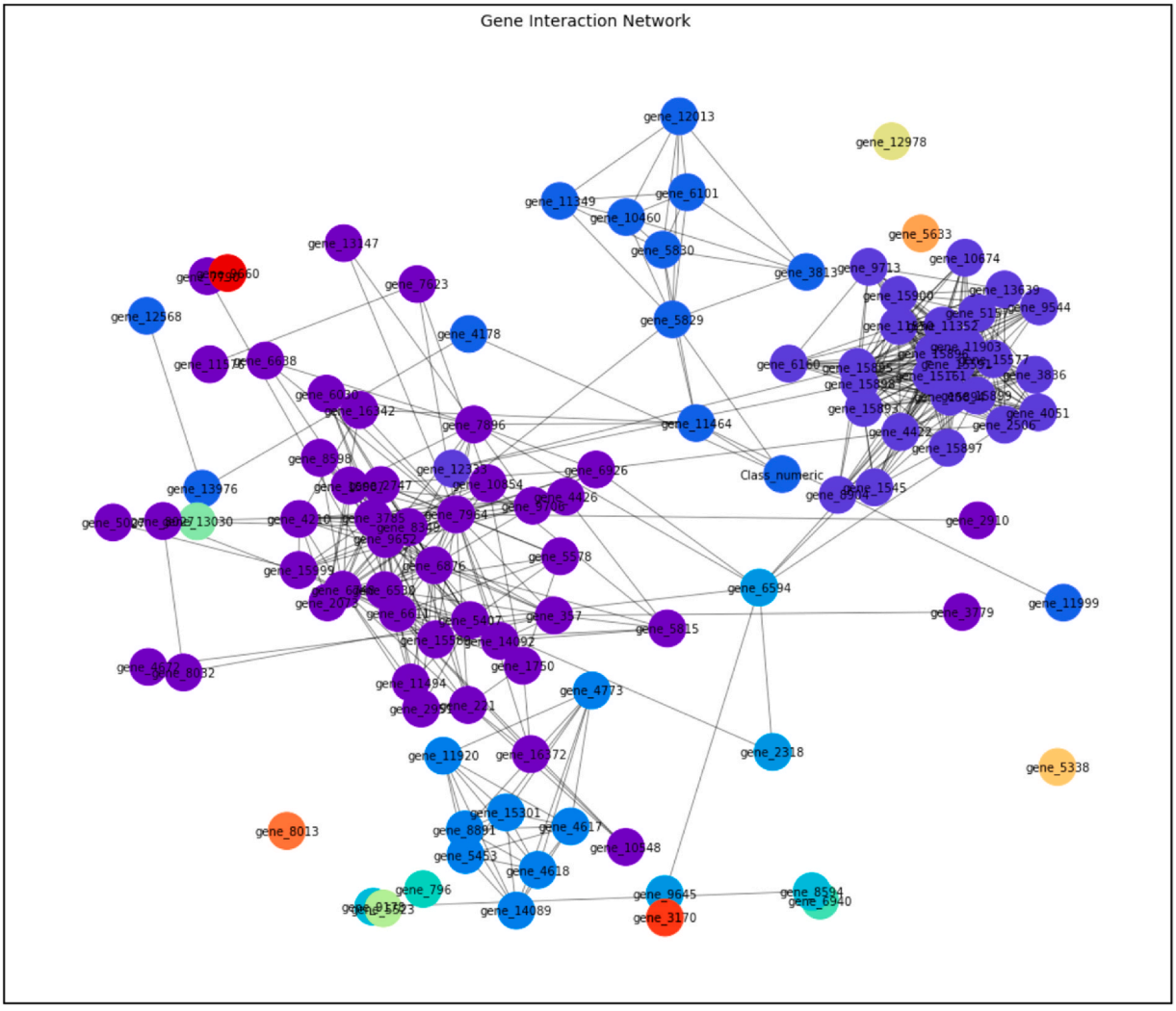


Fig. 6. Gene interaction plot with different communities.

and RF performed well but slightly below SVM. SVM had the best accuracy among all models. After all, the robustness and consistency of SVM in categorizing cancer gene expression profiles are confirmed by this high degree of accuracy.

Fig. 8 compares the outcomes of split-validation and 5-fold cross-validation, with the goal of determining the best robust model for our cancer gene expression data based on selected features. In the split-validation findings, SVM, RF, and ANN all obtained

Table 2
Performance of ML models with split-validation approach.

Models	Accuracy	Precision	Recall	F1 Score	Error Rate
DT	0.8708	0.7138	0.7727	0.7311	0.1291
SVM	0.9958	0.9955	0.9952	0.9953	0.0041
ADA	0.7625	0.7082	0.7692	0.7205	0.2375
QDA	0.7791	0.9258	0.7255	0.7877	0.2208
KNN (n = 10)	0.9916	0.9907	0.9907	0.9905	0.0083
ANN	0.9958	0.9978	0.9987	0.9964	0.0042
RF	0.9958	0.9978	0.9987	0.9965	0.0042
NB	0.9875	0.9872	0.9872	0.9900	0.0125

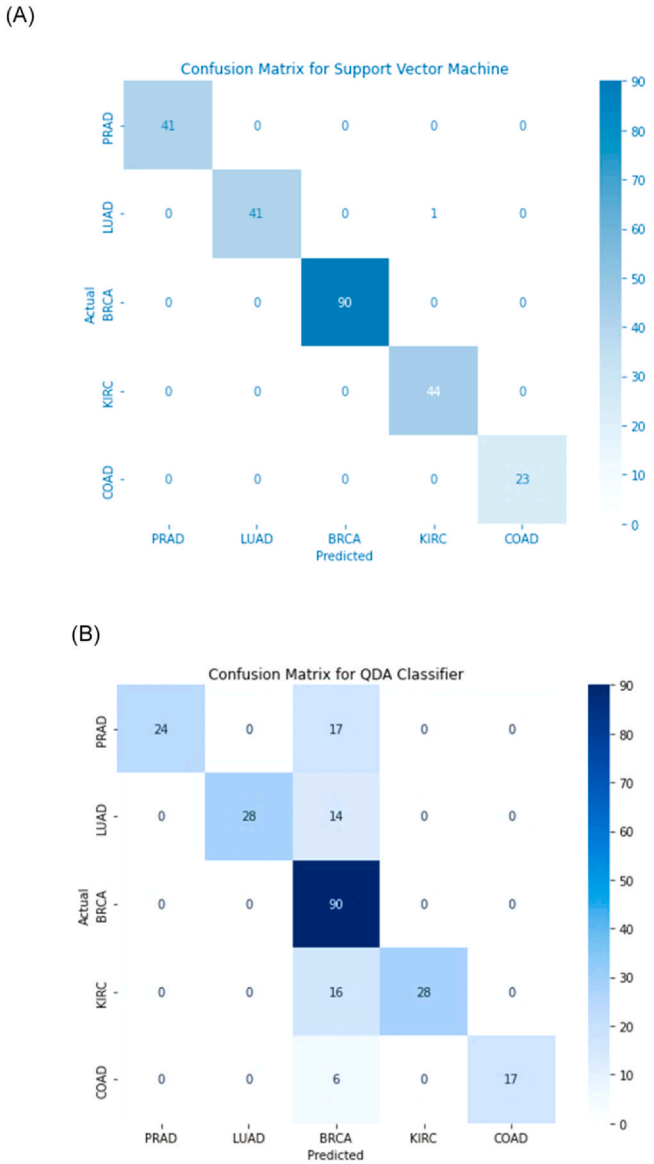


Fig. 7. Confusion Matrices of the (A) SVM and (B) QDA.

99.58 % accuracy with the same error rate of 0.41 %. The KNN model (with k = 10) produced equivalent results. To improve our model selection, we used 5-fold cross-validation to tune the hyperparameters. With the maximum accuracy of 99.87 % and the lowest error rate of 0.13 %, SVM is the most robust model, according to the results, which are displayed in Table 3 and Fig. 8. Cross-validation proved that SVM to be the best option for this dataset after tuning, since no other model could match its performance.

Table 3
Performance of models with 5-fold cross-validation approach.

Models	Accuracy	Error Rate
DT	0.8814	0.1186
SVM	0.9987	0.0013
ADA	0.9312	0.0688
QDA	0.4981	0.5019
KNN	0.9975	0.0025
ANN	0.9912	0.0088
RF	0.9951	0.0049
NB	0.7764	0.2236

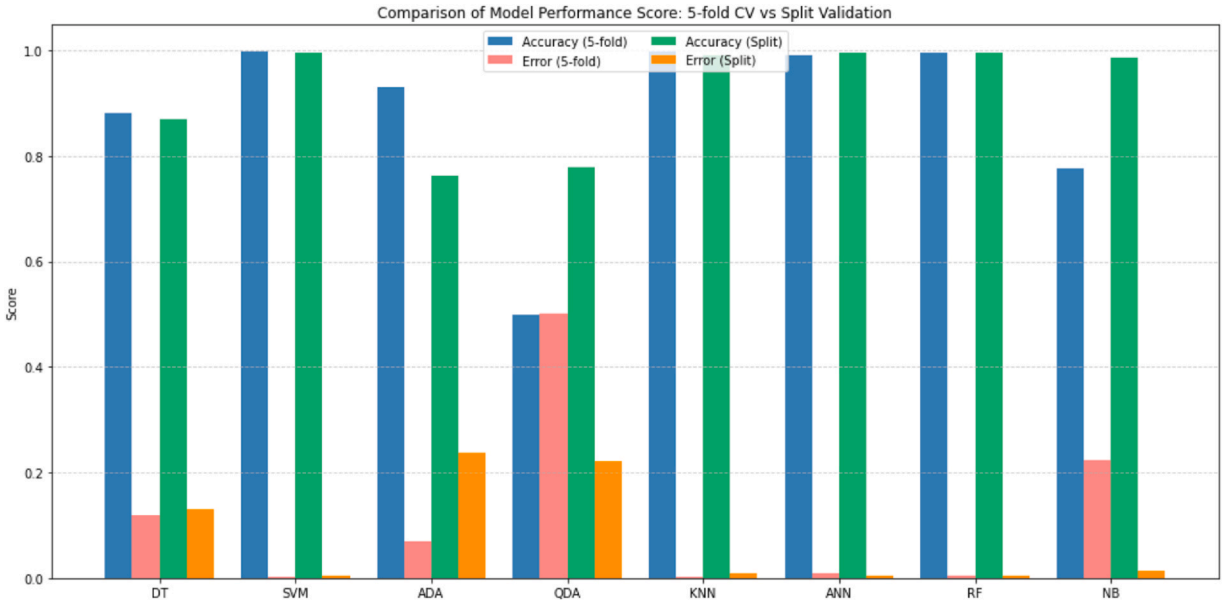


Fig. 8. Summary of performance of models with 5-fold cross-validation and split-validation.

Fig. 9 provides a detailed evaluation of the SVM model, which was identified as the best-performing classifier in this study. In Fig. 9A, a pie chart illustrates the distribution of key performance metrics (Accuracy, Precision, Recall, F1 Score, and Error rate), each contributing approximately equally, except for the error rate, which remains low at 11 %. Fig. 9B presents a concentric donut chart comparing the inner and outer rings representing split- and 5-fold cross-validation performance, respectively. The model achieved 99.87 % accuracy in 5-fold cross-validation with an error rate of just 0.13 %, and 99.58 % accuracy in split-validation with 0.42 % error, demonstrating exceptional reliability and generalizability. To ensure robust performance across all tested models, the SVM model was configured with a linear kernel using parameters C = 1.0 C = 1.0 C = 1.0 and γ = 'scale'\gamma = \text{'scale'}γ = 'scale', which contributed to its superior classification accuracy.

The performance of SVM as the best ML model to classify cancer types in our study was also evident and consistent with other previous studies that tested the model for breast cancer classification [9,10]. Wu and Hicks [9] successfully used SVM to classify triple-negative breast cancers (TNBC) and non-TNBC from gene expression data with high accuracy among three other models (K-nearest neighbor, Naïve Bayes, and Decision tree).

For external validation, we performed our models on an independent dataset (CuMiDa) to ensure robustness and generalizability. The models demonstrated a strong performance as seen in our analysis (Table 4).

Our results collectively highlight the robustness, consistency, and suitability of the SVM model for high-dimensional gene expression data analysis in cancer classification.

Limitations

Despite the importance of our study in cancer diagnosis and control, we recognized a limitation. The dataset we utilized lacked sufficient metadata, such as gene names or descriptions, which made it challenging to identify and understand the biological functions of the genes included in our study.

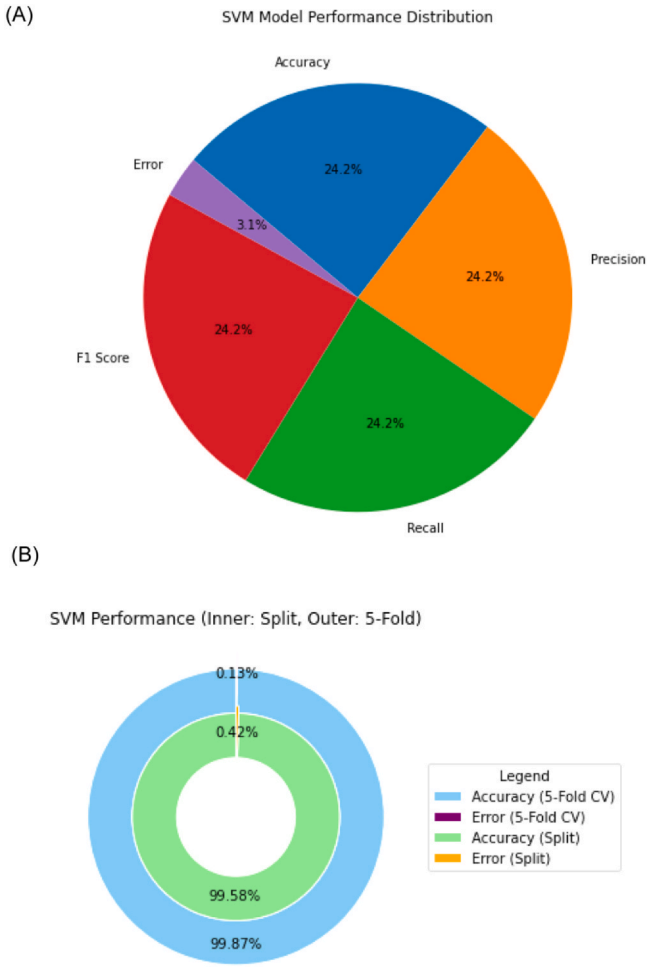


Fig. 9. Distribution of SVM (A) performance and (B) validation on cancer types classification.

Table 4
Performance of models with split- and 5-fold cross-validation approach on CuMiDa.

	Split-validation		5-fold cross-validation	
Models	**Accuracy**	**Error Rate**	**Accuracy**	**Error Rate**
DT	0.58	0.41	0.738	0.262
SVM	0.99	0.01	0.962	0.038
Bagging	0.88	0.12	0.915	0.085
ADA	0.35	0.64	0.730	0.270
GB	0.99	0.00	0.824	0.176
QDA	0.72	0.28	0.438	0.562
KNN	0.92	0.07	0.854	0.146
ANN	0.56	0.43	0.276	0.724
RF	0.99	0.01	0.907	0.093
NB	0.99	0.01	0.862	0.138

Conclusion

The study's findings highlight the effort invested in analyzing gene expression data to improve cancer classification. A refined list of 106 genes associated with various tumor types was identified. Using statistical methods such as Lasso and Ridge regression, the most relevant features were carefully selected. This targeted approach enhances both the analysis and identification of key genetic markers linked to specific cancer types. The performance results of the machine learning models reveal a clear trend. Models like

SVM, KNN, and ANN performed exceptionally well, achieving high accuracy with minimal error in both 5-fold cross-validation and split-validation. This suggests that the dataset may have well-separated class boundaries that these models can effectively capture. In contrast, although Random Forest also showed strong performance, its slightly elevated error rates may hint at overfitting. Among all the models evaluated, Support Vector Machine (SVM) emerged as the best-performing model. It achieved the highest accuracy of 99.87 % with an error rate of just 0.13 % during 5-fold cross-validation. In split-validation, SVM maintained its superior performance with an accuracy of 99.58 % and a minimal error of 0.0041 %. These outstanding results highlight SVM's ability to effectively capture the underlying structure of the dataset, making it a highly reliable model for breast cancer classification. These results prompt further discussion on optimizing model efficiency, ensuring generalizability across diverse data types, and balancing the trade-off between model complexity and interpretability. In the future, a user-friendly graphical interface will be developed for easier access and interaction with the model outputs.

CRediT authorship contribution statement

SA, ROA, and SB conceptualized the idea; SA performed the analysis; SA, ROA, and SB wrote, reviewed, and edited the initial and final draft. All authors agreed on the final draft to be submitted.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability

All the data and code used in this study are available on GitHub: https://github.com/tonikazic/s25_cptnl_gen/blob/master/shreya_et_al/Shreya%2CSultana%20and%20Ridwan/Project_python_script_breast_cancer.py

Declaration of Competing Interest

The authors declare that they have no financial or personal relationship(s) that may have inappropriately influenced them in writing this article.

Acknowledgements

We thank Prof Toni Kazic for her support during the project, supervising and reviewing the project. This manuscript was adapted from our computational genomics final class project.

References

[1] National Cancer Institute. (2024). Cancer statistics. Available at: 〈https://www.cancer.gov/about-cancer/understanding/statistics#:~:text=Cancer%20is%20among%20the%20leading,million%20cancer%2Drelated%20deaths%20worldwide〉.
[2] World Health Organization. (2024). Available at: Global cancer burden growing, amidst mounting need for services. 〈https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing–amidst-mounting-need-for-services〉.
[3] J.T. Loud, J. Murphy, Cancer screening and early detection in the 21st century, Semin. Oncol. Nurs. 33 (2) (2017) 121–128, https://doi.org/10.1016/j.soncn.2017.02.002.
[4] M. Shahzad, M. Rafi, W. Alhalabi, N. Minaz Ali, M.S. Anwar, S. Jamal, M. Barket Ali, F.A. Alqurashi, Classification of clinically actionable genetic mutations in cancer patients, Front. Mol. Biosci. 10 (2024) 1277862, https://doi.org/10.3389/fmolb.2023.1277862.
[5] A.T. Aborode, O.A. Emmanuel, I.A. Onifade, E. Olotu, O.J. Otorkpa, Q. Mehmood, R.O. Adesola, The role of machine learning in discovering biomarkers and predicting treatment strategies for neurodegenerative diseases: a narrative review, NeuroMarkers 2 (1) (2025) 100034.
[6] H.H. Rashidi, J. Pantanowitz, M.G. Hanna, A.P. Tafti, P. Sanghani, A. Buchinsky, B. Fennell, M. Deebajah, S. Wheeler, T. Pearce, I. Abukhiran, S. Robertson, O. Palmer, M. Gur, N.K. Tran, L. Pantanowitz, Introduction to artificial intelligence and machine learning in pathology and medicine: generative and non-generative artificial intelligence basics, Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc. 38 (4) (2025) 100688, https://doi.org/10.1016/j.modpat.2024.100688.
[7] J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, Cancer Inform. 2 (2007) 59–77.
[8] UCI Machine Learning Repository. (n.d.). Gene Expression Cancer RNA-Seq dataset.
[9] J. Wu, C. Hicks, Breast cancer type classification using machine learning, J. Pers. Med. 11 (2) (2021) 61, https://doi.org/10.3390/jpm11020061.
[10] M.F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, Expert Syst. Appl. 36 (2) (2009) 3240–3247.