International Journal for Research in Bioscience
Published by: Green Publications
Article Type: **Research**
Volume 01 Issue 01 March 2026

MACHINE LEARNING-BASED CLASSIFICATION OF BREAST CANCER USING GENE EXPRESSION PROFILES

Riya Pareek[1], Sandeep Kumar[2], Anjali Gupta[3]

[1] Department of Biochemistry, School of Life Sciences, Central University of Rajasthan, India
[2] Department of Genetics, Institute of Life Sciences, India
[3] Department of Bioinformatics, Research Institute of Biotechnology, India

***Corresponding Author:** Riya Pareek*
Email: riyapareek1234@gmail.com

Received:- 17/Dec/2025, Revised:- 12/Jan/2026, Accepted:- 22/Feb/2026, Published:- 24/Mar/2026

Abstract
Breast cancer remains one of the leading causes of cancer-related mortality worldwide, highlighting the need for accurate diagnostic approaches and improved understanding of its molecular mechanisms. Advances in transcriptomic technologies have enabled large-scale analysis of gene expression profiles, providing valuable opportunities for identifying molecular biomarkers associated with cancer development. In this study, gene expression data were analyzed to identify significant genes differentiating tumor and normal breast tissue samples and to evaluate the effectiveness of machine learning models for breast cancer classification. The dataset consisted of 590 samples, including 529 tumor and 61 normal samples, with expression values measured for 17,814 genes. Differential gene expression analysis using a two-sample t-test was performed to identify informative genes, and the top 500 statistically significant genes were selected as predictive features. Three machine learning models Logistic Regression, Support Vector Machine (SVM), and Random Forest were developed to classify tumor and normal samples based on the selected gene expression features. The dataset was divided into training and testing subsets using stratified sampling, and model performance was evaluated using accuracy and receiver operating characteristic area under the curve (ROC–AUC). The results demonstrated strong classification performance, with Logistic Regression and SVM achieving an accuracy of 97.46% and an ROC–AUC of 0.997, while Random Forest achieved an accuracy of 96.61% and an ROC–AUC of 0.994. These findings highlight the potential of combining gene expression analysis with machine learning techniques for breast cancer classification and biomarker discovery.

Keywords: *Breast cancer, Gene expression analysis, Machine learning, Support Vector Machine, Biomarker identification*

1. Introduction

Breast cancer is a common cancer among women in all parts of the world and has been a significant issue in the public health concerning its tendency because of its high rates of occurrence as well as mortality. The global statistics on cancer show that breast cancer is a major percentage of newly diagnosed cases of cancer annually [1]. Early and precise diagnosis is an essential factor in enhancing survival of the patient and proper administration of treatment strategies. The conventional diagnostic methods such as imaging and histopathological methods have been extensively applied in the detection of breast cancer, but however, these modalities might not be effective in identifying the molecular heterogeneity of tumors [2-3].

The advancement of tools for high-throughput sequencing has enabled to analyze in depth the pattern of gene expression that is related to cancer development and progression. Transcriptomic profiling has been developed as an effective method of detecting changes in the molecules and possible biomarkers that contribute to tumorigenesis [4]. The datasets of gene expression give useful data about the activity of thousands of genes at the same time, and researchers can examine intricate biological processes underlining cancer [5]. It has been established by several studies that the tumor tissues usually have different gene expression profiles when compared to the normal ones, an indicator that there is dysregulation of essential cellular processes like cell proliferation, apoptosis and signal transduction [6-7]. What is more, new developments in single-cell and spatial transcriptomics have pointed to the high heterogeneity of breast tumors, further informing the evolution of tumors, their interaction with the microenvironment, and their response to therapy [8]. Although large-scale transcriptomic datasets are available, the analysis of high-dimensional gene expression data is a very challenging task [9]. The quantity of genes easily outnumber the quantity of samples, a factor that may hinder the conventional statistic analysis as well as enhance the chances of overfitting of predictive models. Machine learning methods have thus found greater importance in deriving useful patterns of high-dimensional biological data [10-11]. Along with promising outcomes in cancer classification, biomarker discovery and disease prediction, these computational methodologies can track the discovery of intricate relationships in datasets [12].

Gene expression data has been used in several machine learning algorithms to diagnose cancer, such as “Learning regression”, “Support Vector machine (SVM)” and “the random forest algorithm”. These are especially effective in processing high-dimensional data and determining which features give the best contribution to the performance of classification [13]. Random Forest has also found extensive application in genomic research because it is capable of operating in high-dimensional feature space, whereas Support Vector Machine allows high-dimensional feature space operations but does not have the ability to identify important predictive features via feature importance analysis. Despite the fact that numerous research have employed machine learning techniques to analyze cancer genomics, the study should be furthered to enhance the identification of predictive gene signatures and to assess the predictive ability of the various algorithms on the same using well-defined gene features. The combination of statistical gene selection with Additionally, machine learning algorithms can be utilized to find biologically important genes linked to breast cancer and improve classification accuracy. The objectives of this study were:

1. To identify significantly differentially expressed genes between tumor and normal breast tissue samples using statistical analysis of gene expression data.
2. To develop and evaluate machine learning models, including Logistic Regression, Support Vector Machine, and Random Forest, for classification of breast cancer based on selected gene expression features.
3. To identify key predictive genes contributing to classification performance through feature importance analysis.

2. Materials and Methods

2.1 Dataset Description

The data presented on gene expression in this paper were taken out of a publicly available breast cancer transcriptomic dataset based on "The Cancer Genome Atlas (TCGA)". The data used was a set of 590 breast tissue samples, 529 tumor samples, and 61 non-tumor samples, where the level of expression of 17,814 genes was measured [14]. The dataset was in the form of a row in which the gene was represented and columns that were used to give the values of their expression.

2.2 Data Preprocessing

Before analysis, the dataset experienced a series of preprocessing procedures to ensure data quality and consistency. Missing values in the data were treated with the aid of gene-wise median imputation where missing values of the expression are replaced with the median expression of the respective genes across the samples. The gene expression matrix was imputed and subsequently transposed to put the data in an order that could be subjected to machine learning analysis where samples were in rows and gene features in columns. Every sample was labeled with a class label which was 0 normal tissue and 1 tumor.

2.3 Differential Gene Expression Analysis

Different expression analysis was conducted based on independent two-sample t-test to detect genes that have significant differences in their expressions between tumor and normal specimens. The tumor and normal groups were compared in terms of their respective levels of the expression of each gene and a p-value was calculated. The genes were classified according to their statistical significance and the top 500 genes with the lowest p-values were identified as informative features to be used later as machine learning models. This reduced the number of dimensions and retained the most discriminative genes that were linked to breast cancer.

2.4 Feature Scaling and Data Partitioning

The dataset with the chosen gene features was divided into subgroups for testing and training prior to model training. The stratified sampling method was employed to maintain the proportion of classes of tumor and normal samples. Dividing the data was done in 80 percent training data and 20 percent testing data. In order to produce the appropriate scaling of algorithms that are sensitive to feature magnitude, the StandardScaler method was used to scale gene expression values, which fits features to the mean of zero and unit variance. The training data was used to determine the scaling parameters, which were then projected onto the testing data.

2.5 Machine Learning Models

Based on the chosen gene expression characteristics, three supervised machine learning algorithms were used to categorize tumor and normal data:

Logistic Regression

The Logistic Regression was used as a control model of classification. The algorithm computes the likelihood of membership in classes by modeling the probability as a logistic function and approximates the model parameters by the use of maximum likelihood optimization.

Support Vector Machine

"Support Vector Machine (SVM) classification was performed using a linear kernel, which is commonly applied in high-dimensional genomic datasets". The algorithm identifies an optimal hyperplane that maximizes the separation margin between classes in the feature space.

Random Forest

Classification was also done using an ensemble learning technique, which is the random forest, and is based on the decision trees. The model was built in 200 decision trees and the majority voting

mechanism between decision trees was used to give predictions. Besides classification, Random Forest was employed to calculate feature importance scores and, therefore, to identify the genes that are the most significant with respect to classification.

2.6 Model Evaluation

The machine learning models' performance was evaluated using the independent testing dataset. Various metrics of evaluation were used in the assessment of classification performance such as accuracy, precision, recall, F1-score, and the area under receiver operating characteristic curve (ROC–AUC). These measures gave the holistic evaluation of the predictive quality of each model in differentiating tumor and normal samples. To summarize the classification results, and determine the correctly and incorrectly classified samples of the classes, confusion matrices were created. Moreover, receiver operating characteristic (ROC) curves were also plotted to represent the sensitivity-specificity trade-off of the various classification thresholds to facilitate a comparison between the discriminative ability of the developed models.

2.7 Visualization and Data Analysis

To further investigate the form of the data of gene expression and demonstrate the results of the classification, a number of visualization methods were used. To reduce the size of the selected gene characteristics and show the difference between the tumor and normal samples on a two-dimensional plane, “Principal Component Analysis (PCA)” was used. To determine the distribution of differentially expressed genes in relation to the variation in expression and the statistical significance of the variation, the data was shown in a volcano plot. A heatmap of the 20 highest-ranking predictive genes provided by the “Random Forest model” was also generated to demonstrate the patterns of expression in the samples. All the analyses and visualizations were done with Python programming language and libraries such as the pandas, NumPy, scikit-learn, matplotlib, seaborn, and SciPy.

3. Results

3.1 Dataset Characteristics and Preprocessing

The gene expression dataset comprised 590 breast tissue samples, of which 529 were tumor samples and 61 were normal samples, which were measured on 17,814 genes. The data set contained missing values which were identified and filled using gene-wise median imputation leading to full data set to be analyzed in the downstream.

Differential expression analysis was conducted on an independent t-test of two samples to compare tumor and normal samples to reduce the dimensions and pay attention to informative features. The statistical significance was used to rank the genes, and the 500 genes having the lowest p-values were taken to a machine learning analysis.

3.2 Differential Gene Expression Analysis

Differential expression analysis revealed that certain genes were more highly expressed in the tumor samples than in the normal samples. Some of the strongest differences in the level of expression were the ITM2A, SFRP1, OSR1, HOXA4, and UBE2E3, which were among the most statistically significant genes. The volcano plot depicting the correlation between differences in the expression and statistical significance of the results is shown in Figure 1 and it portrays the distribution of differentially expressed genes. Genes having large differences in their expression and a low p-value are found towards the top part of the plot. Genes on the right of the plot are genes that are expressed in a greater proportion in tumor cells as compared to those on the left which are genes which are expressed in less proportions as compared to normal cells.

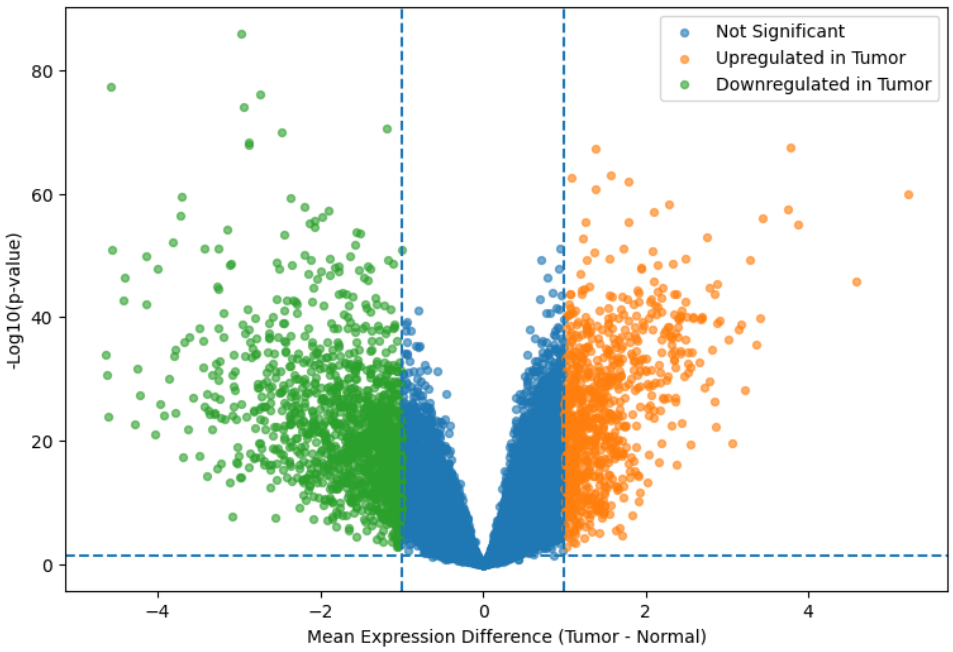


Figure 1: Volcano plot of differentially expressed genes between tumor and normal breast tissue samples.

3.3 Principal Component Analysis

The selected gene expression features were subjected to "Principal Component Analysis (PCA)" in order to examine the overall structure of the data set. PCA was used to indicate that the tumor and normal samples were well separated along the first principal component as indicated in Figure 2. Normal samples were relatively compactly clustered, except that the tumor samples were more widely dispersed in the major component space. This disjuncture suggests that the chosen features of gene expression do note a significant difference between the two sample groups.

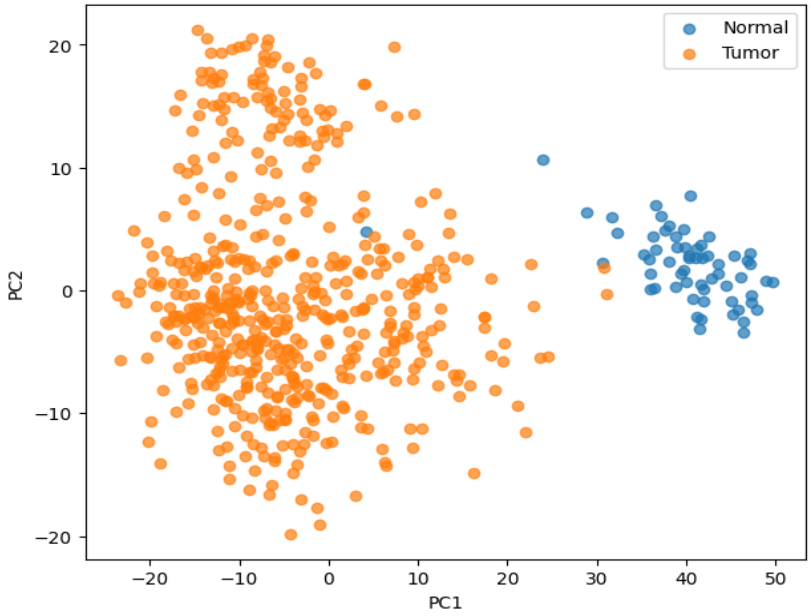


Figure 2: Principal component analysis of gene expression profiles in tumor and normal breast tissue samples.

3.4 Machine Learning Model Performance

To classify tumor and normal samples using the chosen gene expression features, "three supervised machine learning algorithms" including the "Logistic Regression, Supportive Vector machine (SVM),

and Random Forest" were implemented. After the data was split into training and testing subsets, 80% of it was used to train the model and 20% was utilized for testing. Stratified sampling was used to preserve the initial distribution of classes of both tumor and normal samples in two subsets.
The input features that were chosen were the gene features in which the models were trained and then tested against the independent testing features. Model performance was evaluated using classification accuracy and the area under the receiver operating characteristic curve (ROC-AUC). Table 1 summarizes the performance of the three models.

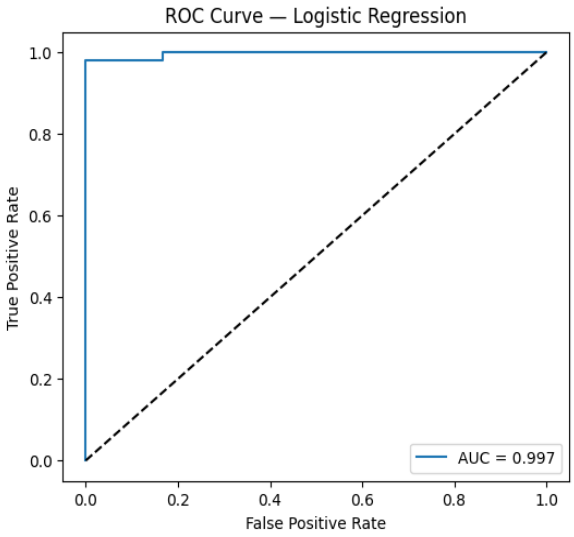
Table 1. Performance metrics of the three models

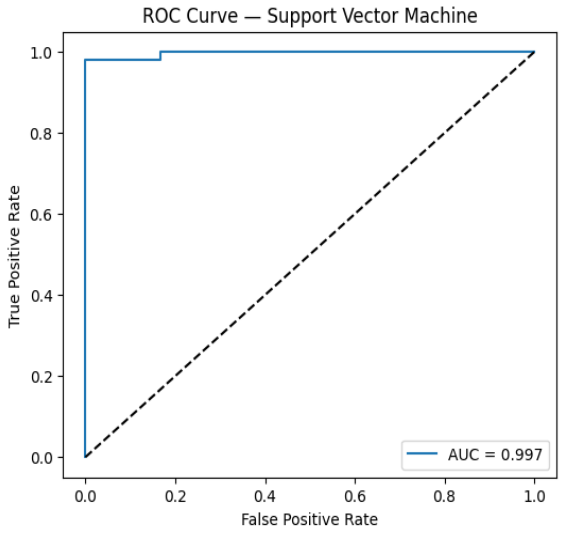
Model	Accuracy	ROC-AUC
Logistic Regression	0.9746	0.9969
Support Vector Machine	0.9746	0.9969
Random Forest	0.9661	0.9937

Both Logistic Regression and "Support Vector Machine" achieved the highest classification accuracy of 97.46%, indicating strong predictive performance in distinguishing tumor and normal samples. The Random Forest model also demonstrated high classification accuracy, achieving 96.61%. All models achieved ROC–AUC values greater than 0.99, indicating excellent discriminative capability across the evaluated classifiers.

3.5 Receiver Operating Characteristic Analysis

"Receiver Operating Characteristic (ROC) curves were created to further discuss the performance of the machine learning models in terms of classification.The trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) at different classification thresholds is shown graphically by ROC curves. The results of ROC of Logistic Regression, Support Vector Machine as well as the random Forest are represented by Figure 2, respectively. "A quantitative measure of model performance was taken to be the area under the ROC curve (AUC)."
Both the Logistic Regression and SVM generated almost the same ROC curves whose AUC is 0.997, which means that classification performance is extremely high and there is a strong separation between the tumor and normal samples. Random Forest model attained an AUC of 0.994 which also depicts a good predictive ability. In general, the ROC analysis shows that all three machine learning models have high classification performance in case they are trained on the chosen gene expression features.





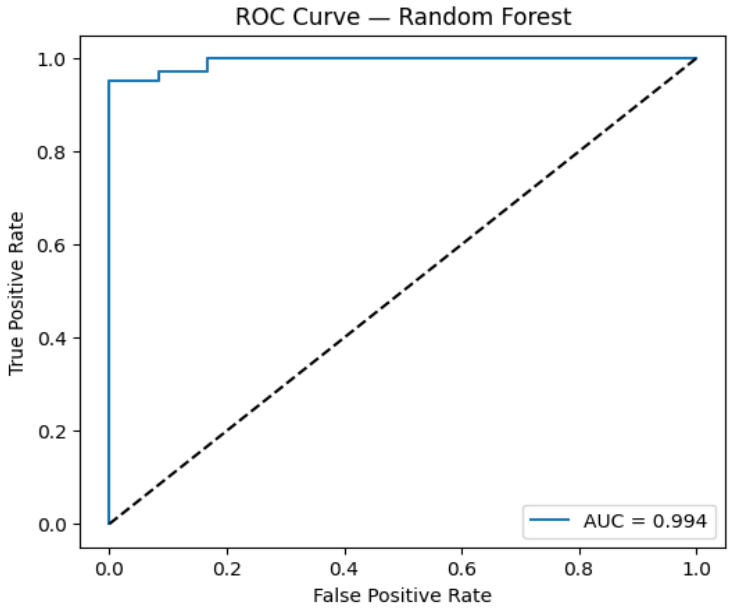


Figure 2: Receiver operating characteristic (ROC) curve (a. Logistic Regression classification model, b. Support Vector Machine, c. Random Forest)

3.6 Identification of Predictive Genes

"The Random Forest model" was used to determine the importance of the features by using it to analyse the significance of the genes to the classification performance. Table 2 gives the top 20 most important predictive genes. Among the most influential genes were found TSLP, ADAMTS5, SPRY2, SCN4B, FIGF, MMP11, CDCA8, MME, PAQR4 and IGSF10, among others, in order to test the effect of these genes on the classification performance, feature importance analysis on the Random Forest model was performed. Random Forest also has importance scores of each feature according to their role in the decision-making process throughout the group of decision trees.

Table 2 shows the top 20 most important predictive genes in the ranking of the genes. The most important genes included TSLP with the score of importance calculated as 0.0969, then ADAMTS5 with the score of importance calculated as 0.0464 and finally SPRY2 with the score of importance calculated as 0.0395. The other genes with high ranking were SCN4B, FIGF, MMP11, CDCA8, MME, PAQR4, and IGSF10, which showed that they were strongly weighted on the classification models. These genes are the most descriptive in the dataset and they were also very important in differentiating the normal and tumor samples in the machine learning analysis.

Table 2. Top predictive genes identified by Random Forest feature importance

Rank	Gene	Importance Score
1	TSLP	0.0969
2	ADAMTS5	0.0464
3	SPRY2	0.0395
4	SCN4B	0.0297
5	FIGF	0.0291
6	MMP11	0.0290
7	CDCA8	0.0274
8	MME	0.0258
9	PAQR4	0.0257
10	IGSF10	0.0213
11	GPRIN1	0.0211
12	COL10A1	0.0197
13	SDPR	0.0184

14	H2AFX	0.0173
15	CA4	0.0172
16	KLHL29	0.0168
17	TPX2	0.0164
18	MAMDC2	0.0151
19	FXYD1	0.0148
20	FREM1	0.0143

3.7 Expression Patterns of Predictive Genes

A heatmap visualization of the top 20 most informative genes as indicated by the Random Forest feature importance analysis was created to further examine expression profiles of the most informative genes. The heatmap provided in Figure 3 depicts the patterns of gene expression in all samples in the dataset. The heatmap represents the genes in each row and a sample in each column. The color value is used to show the relative expression level of the gene, which can be used to see the differences in expression of both tumor and normal samples. There is a variety of genes being active in the selected dataset, and the distinct patterns of their expression can be observed. This visualization emphasizes the top important predictive gene expression as well as it gives a summary of transcriptional variation among the samples under analysis.

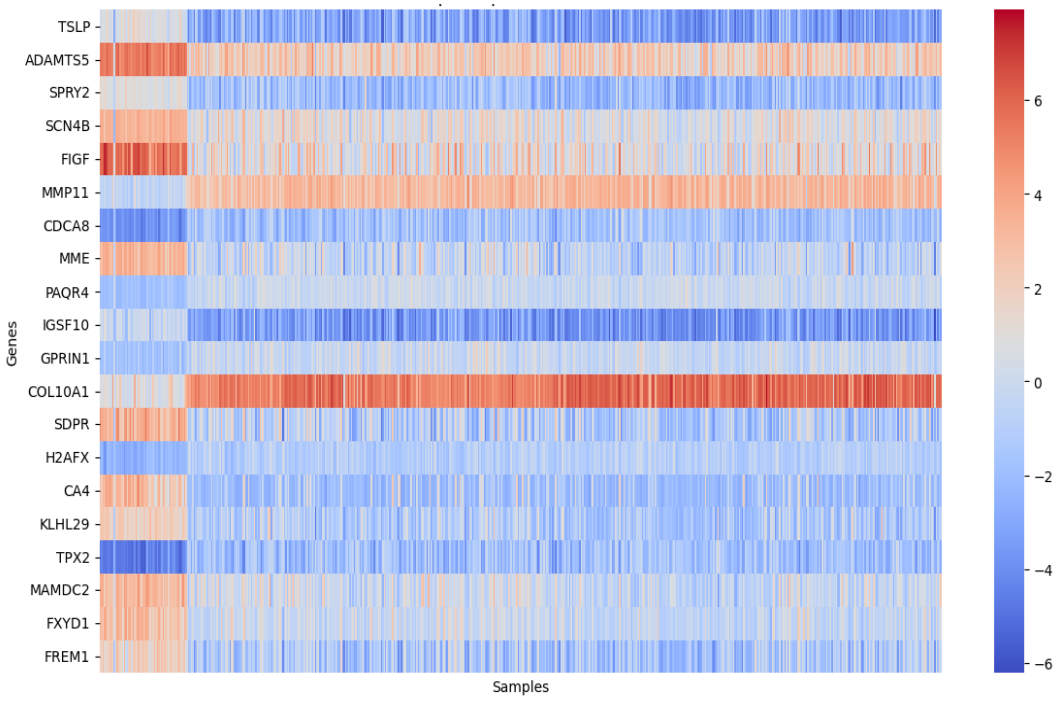


Figure 3: Heatmap showing expression patterns of the top 20 predictive genes identified by the Random Forest model.

4. Discussion

"One of the most prevalent types of cancer worldwide is breast cancer and early detection is necessary to enhance the outcomes of patients." Transcriptomic technologies have had significant impact due to their ability to analyze gene expression profiles in large scale to give important insights into the molecular mechanisms involved in cancer development [15]. Gene expression data of this research were processed so as to predict the presence of predictive gene signatures and the efficacy of machine learning models in differentiating between tumor and normal breast tissue samples.

The analysis of difference in gene expression showed a great number of genes with significant differences in their expressions between the tumor and normal samples. The volcano plot showed that there is a general transcriptional dysregulation in the cancer tissues of the breast, which indicates the

vast molecular alterations in tumorigenesis. These transcriptional modifications are commonly observed in cancer biology wherein perturbed gene expression dumpsters tend to indicate perturbation of signaling pathways, altered cell cycle control, and cellular metabolic modifications [16- 17]. The discovery of the markedly dysregulated genes hence offers a valuable foundation within the designing of predictive biomarkers that are used to detect cancer. Principal component analysis also indicated that the expression patterns of genes are significantly different in tumor and normal samples. The PCA plot indicated a separation between the two groups with the features of the most commonly used genes incorporating high levels of biological variability relative to breast cancer [18]. There was more variation in the PCA space whether a tumor sample or normal sample was given, and the finding indicates that transcriptional heterogeneity is higher in tumor tissues. This heterogeneity is typical of cancer data and represents the intricate molecular topography of tumor evolution.

The machine learning models trained on this paper have reached high classification accuracy on the selected gene expression features. Both the Logistic Regression and Support Vector Machine models had high levels of classification and ROC–AUC of about 1.0 which denotes good discrimination between tumor and normal samples. Random Forest also showed a high level of predictive performance as the accuracy and ROC-AUC value were slightly lower [19]. These findings bring out the usefulness of machine learning methods in the analysis of high-dimensional transcriptomic data and in the identification of molecular patterns that can be related to disease conditions. It indicates the good results of Logistic Regression and SVM models that the gene expression features chosen give the clear separation of tumor and normal samples. Earlier research has also discovered that linear classifiers are useful in helping to analyze high-dimensional gene expression data in case informative genes are identified by statistical filtering [20]. High values of ROC-AUC in this research means that the gene expression signatures do have adequate information that can assist in the correct classification of the breast cancer samples as is the case with the gene expression-based methods of classification that have been used before [21].

The analysis of feature importance with the help of the Random Forest model has selected some genes which made a significant contribution to the classification performance. Most of the most influential genes included TSLP, ADAMTS5, SPRY2, SCN4B, FIGF, MMP11, CDCA8, MME and PAQR4. Some of these genes have been previously reported to be involved in biological processes of cancer. As an example, MMP11 is a part of the family of matrix metalloproteinase, which is considered significant in extracellular matrix remodelling and tumor invasion [22]. An elevation in the expression of MMP11 has also been linked to higher levels in tumor cell migration and worse clinical prognosis in breast cancer [23- 24). Likewise, SPRY2 takes part in the cell proliferation and differentiation signaling pathways and its disregulation has been associated with breast cancer metastasis. These results indicate that the machine learning models learned biologically important transcriptional patterns that are related to breast cancer.

The heatmap analysis of the most predictive genes also showed a clear difference in the expression of the genes in the samples indicating that the genes are useful in the distinction of tumor and normal tissues. These signatures of gene expression can be potential molecular markers of breast cancer classification, and can give information on pathogenesis of the disease. Even though the models had high predictive accuracy, there are a number of limitations that ought to be taken into account. First, the analysis was carried out based on one dataset of gene expression and this can be considered as a limitation in the generalizability of results to various cohorts or experimental frameworks. Second, the data available on gene expression might not be sufficient to reflect the complexity of cancer biology because there are other factors of the molecular level, including epigenetic modifications, protein changes, and genomic mutations, which lead to tumor formation. Future research may consider combining the multi-omics data to achieve a more holistic view of the breast cancer molecular mechanisms and enhance the models of prediction. Irrespective of these shortcomings, the results of this study show that transcriptomic analysis, when used in combination with machine learning methods offers a potent methodology to predictive gene signatures and distinguish between tumor and normal samples of the breast tissue. The excellent performance of the models in

classification indicates the usefulness of gene expression-based systems in designing diagnostic applications and promoting precision medicine in the study of breast cancer.

5. Conclusion

The performance of the combination of transcriptomic analysis and machine learning methods in the classification of breast cancer tissue samples using gene expression patterns was shown in this paper. The study of the differential gene expression was conducted that revealed a small number of highly significant genes that differentiated the tumor and the normal breast tissue samples. Three machine learning models, Logistic Regression, Support Vector Machine and Random Forest were built using the top 500 statistically significant genes as predictive feature to classify tumor and normal samples. All models had high predictive performance, but the highest accuracy and ROC-AUC values were obtained by Logistic Regression and Support Vector Machine. Such findings suggest that the signatures of gene expression present highly informative characteristics of the differentiation of breast cancer tissue and normal samples. Moreover, the feature importance analysis revealed that TSLP, ADAMTS5, SPRY2, SCN4B, FIGF and MMP11 were some of the key features which helped in classifying them. On balance, the results demonstrate how gene expression profiling with machine learning solutions can be used to classify and discover biomarkers to accurately classify and uncover the root causes of breast cancer. Future research can build upon this by confirming the gene signatures as identified above in independent cohorts and by combining other molecular data to improve predictive models and better insight into the biology of breast cancer.

References

1. Giaquinto, A. N., Sung, H., Newman, L. A., Freedman, R. A., Smith, R. A., Star, J., ... & Siegel, R. L. (2024). Breast cancer statistics 2024. *CA: a cancer journal for clinicians*, *74*(6), 477-495.
2. Hacking, S. M., Yakirevich, E., & Wang, Y. (2022). From immunohistochemistry to new digital ecosystems: a state-of-the-art biomarker review for precision breast cancer medicine. *Cancers*, *14*(14), 3469.
3. Lopez-Gonzalez, L., Sanchez Cendra, A., Sanchez Cendra, C., Roberts Cervantes, E. D., Espinosa, J. C., Pekarek, T., ... & Diaz-Pedrero, R. (2024). Exploring biomarkers in breast cancer: hallmarks of diagnosis, treatment, and follow-up in clinical practice. *Medicina*, *60*(1), 168.
4. Albitar, M., Goy, A., Pecora, A., Graham, D., McNamara, D., Charifa, A., ... & Waintraub, S. (2024). The use of transcriptomic data in developing biomarkers in breast cancer. *ImmunoMedicine*, *4*(1), e1051.
5. Kawiak, A. (2022). Molecular research and treatment of breast cancer. *International journal of molecular sciences*, *23*(17), 9617.
6. Moar, K., Pant, A., Saini, V., Pandey, M., & Maurya, P. K. (2023). Potential diagnostic and prognostic biomarkers for breast cancer: A compiled review. *Pathology-Research and Practice*, *251*, 154893.
7. Zhu, S., Zhang, M., Liu, X., Luo, Q., Zhou, J., Song, M., ... & Liu, J. (2023). Single-cell transcriptomics provide insight into metastasis-related subsets of breast cancer. *Breast Cancer Research*, *25*(1), 126.
8. Han, X., Li, X., Bai, L., & Zhang, G. (2025). Single-cell transcriptomics in metastatic breast cancer: mapping tumor evolution and therapeutic resistance. *Frontiers in Genetics*, *16*, 1669741.
9. Wang, X., Venet, D., Lifrange, F., Larsimont, D., Rediti, M., Stenbeck, L., ... & Sotiriou, C. (2024). Spatial transcriptomics reveals substantial heterogeneity in triple-negative breast cancer with potential clinical implications. *Nature communications*, *15*(1), 10232.
10. An, J., Lu, Y., Chen, Y., Chen, Y., Zhou, Z., Chen, J., ... & Peng, F. (2024). Spatial transcriptomics in breast cancer: providing insight into tumor heterogeneity and promoting individualized therapy. *Frontiers in Immunology*, *15*, 1499301.

11. Zhang, Y., Gong, S., & Liu, X. (2024). Spatial transcriptomics: a new frontier in accurate localization of breast cancer diagnosis and treatment. *Frontiers in Immunology*, *15*, 1483595.
12. Rezaei, S., Hamedani, Z., Ahmadi, K., Ghannadikhosh, P., Motamedi, A., Athari, M., ... & Arabi, H. (2025). Role of machine learning in molecular pathology for breast cancer: A review on gene expression profiling and RNA sequencing application. *Critical Reviews in Oncology/Hematology*, *213*, 104780.
13. Chen, X., Yi, J., Xie, L., Liu, T., Liu, B., & Yan, M. (2024). Integration of transcriptomics and machine learning for insights into breast cancer: exploring lipid metabolism and immune interactions. *Frontiers in Immunology*, *15*, 1470167.
14. Orvile. (2025). *Breast cancer gene expression dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/orvile/gene-expression-profiles-of-breast-cancer
15. Sahu, D., Shi, J., Segura Rueda, I. A., Chatrath, A., & Dutta, A. (2024). Development of a polygenic score predicting drug resistance and patient outcome in breast cancer. *NPJ Precision Oncology*, *8*(1), 219.
16. Thalor, A., Joon, H. K., Singh, G., Roy, S., & Gupta, D. (2022). Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. *Computational and structural biotechnology journal*, *20*, 1618-1631.
17. Mirza, Z., Ansari, M. S., Iqbal, M. S., Ahmad, N., Alganmi, N., Banjar, H., ... & Karim, S. (2023). Identification of novel diagnostic and prognostic gene signature biomarkers for breast cancer using artificial intelligence and machine learning assisted transcriptomics analysis. *Cancers*, *15*(12), 3237.
18. Park, J. W., & Rhee, J. K. (2024). Integrative analysis of ATAC-seq and RNA-seq through machine learning identifies 10 signature genes for breast cancer intrinsic subtypes. *Biology*, *13*(10), 799.
19. Tschodu, D., Lippoldt, J., Gottheil, P., Wegscheider, A. S., Käs, J. A., & Niendorf, A. (2023). Re-evaluation of publicly available gene-expression databases using machine-learning yields a maximum prognostic power in breast cancer. *Scientific Reports*, *13*(1), 16402.
20. Muthamilselvan, S., & Palaniappan, A. (2023). Brcadx: Precise identification of breast cancer from expression data using a minimal set of features. *Frontiers in Bioinformatics*, *3*, 1103493.
21. Di Cosimo, S., Pizzamiglio, S., Ciniselli, C. M., Duroni, V., Cappelletti, V., De Cecco, L., ... & Verderio, P. (2024). A gene expression-based classifier for HER2-low breast cancer. *Scientific Reports*, *14*(1), 2628.
22. Kwon, M. J. (2023). Matrix metalloproteinases as therapeutic targets in breast cancer. *Frontiers in oncology*, *12*, 1108695.
23. Kang, S. U., Cho, S. Y., Jeong, H., Han, J., Chae, H. Y., Yang, H., ... & Kwon, M. J. (2022). Matrix metalloproteinase 11 (MMP11) in macrophages promotes the migration of HER2-positive breast cancer cells and monocyte recruitment through CCL2–CCR2 signaling. *Laboratory investigation*, *102*(4), 376-390.
24. Kim, H. S., Kim, M. G., Min, K. W., Jung, U. S., & Kim, D. H. (2021). High MMP-11 expression associated with low CD8+ T cells decreases the survival rate in patients with breast cancer. *PLoS One*, *16*(5), e0252052.