Predictive Insights: Unmasking Breast Cancer Biomarkers through machine learning and Systems Biology

A A Zainulabidin[1,2], A J Sufyan[3], M K Thirunavukkarasu[3*]

[1] Floret Center for Advance geneomics and Bioinformatic Research, Lagos State, Nigeria
[2] Shaheed Center for Research and Advanced Studies, Plateau State, Nigeria
[3] School of Sciences and Humanities, SR University, Warangal, Talangana – 506371, India

*Corresponding author's email: t.muthukumar1996@gmail.com

Abstract. Breast cancer is a complex and heterogeneous disease in nature with quite high rates of metastasis and recurrence that cause significant morbidity and mortality. Despite the improved treatment options with new medical therapies, a proper understanding of the molecular mechanism in breast cancer development and its progression is of utmost necessity. Hence, we conducted a comprehensive analysis on transcriptomic profiling combined with SHAP feature importance calculation in an attempt to find potential molecular targets. Among the 9 machine learning models generated, random forest model displayed an accuracy value of 0.96 for breast cancer prediction. KRT17, KRT5 and FABP5 were the commonly resulted prognostic biomarkers during the DGE and feature selection approaches. Furthermore, gene enrichment and functional annotations of key genes reveals the importance of these key genes in breast cancer progression. The survival analysis confirms the risk associate with key genes in breast cancer patients. Therefore, this finding show the effectiveness of machine learning combine with DGE in Biomarkers discovery and experimental validation of these genes would be a promising approach to eliminate the clinical complications during the breast cancer treatment.
Keywords: *Breast cancer; Machine learning; DEGs; SHAP; Survival Analysis*

1. Introduction

Breast cancer (BC) often detected and poses a significant risk to the health of women [1]. The majority of women die from cancer due to high rates of metastasis and recurrence, even though many breast cancer patients have benefited from advances in medical therapies [2]. It comprises 22.9% of invasive cancers in women and 16% of all female cancers, and at the molecular level (DNA, RNA, proteins, and the metabolites) which is at the level of molecules that control all the biological events and structures, which together regulate cellular behavior and disease development [3], it is a heterogeneous disease. According to its molecular characteristics, breast cancer can be divided into three types: BRCA mutation, hormone receptor (HR: estrogen receptor and progesterone receptor) activation, and human epidermal growth factor receptor 2 (HER2, encoded by ERBB2) activation [1]. Despite the improved treatment options with new medical therapies. The proper understanding of the molecular mechanism in breast cancer development, is of utmost necessity [3], that is the mechanisms governing breast cancer initiation and progression, which is fundamental to identifying novel biomarkers, understanding tumor heterogeneity, and developing more precise diagnostic and therapeutic strategies. These mechanisms provide insights into cellular signaling pathways, genetic alterations, and transcriptional dysregulation

that drive tumorigenesis and metastasis [1]. Highly potent research methods can be used to identify changes in levels of mRNA gene expression [4]. Using microarray technology for gene expression profiling, studies have identified differentially expressed genes (DEGs) that play a critical role in the occurrence and progression of breast cancer, which also have the potential of becoming drug targets and diagnostic markers [5]. Similarly, the biomarkers associated with BRCA 1 cancers has been identified with the aid of mRNA profiling of breast cancer patients using bioinformatics analysis [6].

Machine learning has become very popular in recent years for biomarker discovery [7]. Kloten et al. (2013) identified ITIH5 and DKK3 promoter methylation as potential biomarkers for breast cancer screening, achieving 41% sensitivity and 93-100% specificity. The combination of these genes with RASSF1A methylation increased sensitivity to 67%, suggesting a potential biomarker for early-stage breast cancer screening [8]. Taghizadeh et al (2022) has predicted the breast cancer biomarkers with transcriptomic profiling by using logistic regression and Multilayer perceptron classifier which resulted an accuracy of 0.86 and AUC of 0.94 [9]. Similarly, Gamble et al. (2021) predicted the biomarkers with the help of deep learning systems using pathological slide images [10]. previous models often suffer from limited sensitivity in early detection, overfitting due to high-dimensional gene expression data, lack of generalizability, and insufficient biological interpretability [3]. Our integrated DGE–ML framework addresses these by performing rigorous feature selection, using multiple algorithms to enhance predictive power, and incorporating biological interpretation of key genes. Investigation of these potential biomarkers by different modeling approaches are essential for validating the research findings.

In this study, we explored potential molecular targets and signaling pathways related to the occurrence and development of breast cancer patients at the genomic level by combining traditional DGE (Differential gene expression) analysis combined with machine learning models. This could offer a crucial theoretical foundation for identifying novel treatment targets for breast cancer, by integrating machine learning algorithms which has most capability and predicting powers[11]. In addition to the DEGs and hub genes found by bioinformatics research, prognostic signatures, functional annotations, potential prognostic value, and protein-protein interaction (PPI) networks have all been thoroughly investigated in this instance. Besides, the results of this study may have useful applications in the progress of personalized medicine and in bettering patient outcomes.

2. Methods

Data set

The dataset GSE229005 which included 323 disease samples and 20 control samples, was retrieved from The National Centre for Biotechnology Information (NCBI) database. The samples were processed using GPL33315 platform (Upregulation of VEGF-Hypoxia Signature in Black Women with Breast Cancer). Initial preprocessing steps such as filling the missing values, log transformation of expression values and remove the noise signals are performed to eliminate the false positive prediction. We have build the models to predict the cancer status (tumor/non-tumor) with the help of gene expression values obtained from patient samples.

Identification of DEGs

The R software version 4.0.0 was used to carry out detailed analyses of the DEGs between breast cancer and normal samples. DESeq2 package was utilized to find the DGEs. Genes with a log FC ≥ 1.5 were considered up-regulated, while those with logFC ≤ -1.5 were considered down-regulated; thus, they gave insight into gene expression changes between cancer and normal samples. Also, we have used statistical t-test analysis to find the significant genes among the DGEs ($p>0.05$).

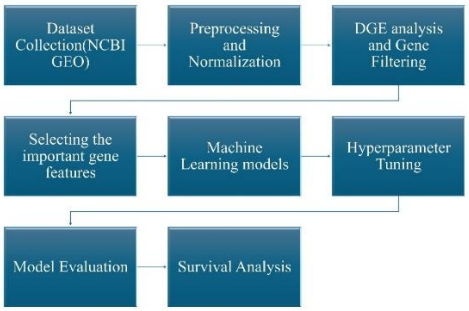


Figure 1. A step-by-step explanation of model construction and hub-gene evaluation

Machine Learning Algorithms

Nine ML algorithms were generated using the dataset. Before model generation, the dataset samples were classified into two such as active and inactive, where the active is the disease samples with activity as 1 and controls are the normal samples as inactive with activity as 0. To create ML models, we used Google Colab Notebook. First, the workspace was filled with the necessary Python packages, including matplotlib, rdkit, sklearn, and others. Subsequently, the following models were accessed to predict the status of breast cancer. Each model we have fine tuned with the help of hyperparameters such as learning rate, max_depth, n_estimators for Random Forest and XGBoost; C and penalty for Logistic Regression. Hyperparameters were optimized using grid search with 5-fold cross-validation to avoid overfitting. Final, models were constructed with the optimized parameters to predict the cancer status effectively.

Random Forest Classifier (RFC)

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using random subsets of the data and aggregates their outputs to enhance prediction accuracy. It is widely used for classification tasks due to its robustness to overfitting and its ability to handle large feature spaces [3].

Logistic Regression (LR)

Logistic Regression is a statistical model that predicts the probability of a binary or multiclass outcome based on one or more independent variables. Despite its simplicity, it remains a powerful and interpretable tool for many biomedical classification problems [12].

Naïve Bayes (NB)

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, assuming strong independence among features. It is particularly efficient for high-dimensional data and often performs well despite its simplistic assumptions [13].

Gradient Boosting (GB)

Gradient Boosting is an ensemble method that constructs additive predictive models in a sequential manner. It minimizes the loss function by fitting new models to the residual errors of previous models, usually using shallow decision trees as base learners [3].

Extreme Gradient Boosting (XGBoost or XGB)

XGBoost is an optimized implementation of gradient boosting that incorporates regularization, parallel processing, and advanced tree pruning strategies. It is highly scalable and has become a leading choice in ML competitions and real-world applications [3].

K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm that classifies data based on the majority label of its 'k' nearest neighbors in the feature space. It is simple to implement and effective for small- to medium-sized datasets where class distributions are locally homogeneous [3].

AdaBoost Classifier
Adaptive Boosting (Ada Boost) combines multiple weak classifiers in sequence, where each subsequent model focuses on the instances that were misclassified by its predecessors. By adjusting the weights of training instances, the model improves performance iteratively until convergence or a minimal error threshold is reached [3].

Cat Boost
CatBoost is a powerful algorithm for gradient boosting that has gained much attention, especially in tasks involving natural language processing. One of the strong points of this algorithm is its very good handling of categorical data, which means that it can automatically perform preprocessing of categorical data and represent it in a way that allows the model to learn better. [3]].

Model prediction evaluation
The ML algorithms that were generated using the gene expression dataset. Each model was validated using external validation metrics to evaluate the predictive capacity of each modal. The evaluation accuracy metrics. Each model was evaluated by the following parameters: Prediction Accuracy is defined as the ability of a model to differentiate actives and inactive cases correctly. Precision and recall score are the two important parameters which is used to evaluate the predictive performance of both positive and negative targets [11]. F1 measure is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test. All the external set validation results for the best performing. The evaluation metrics as follows:

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

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

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

$$F1\ Score = \frac{2*Precision*Recall}{Precision\ +Recall}$$

SHAP feature importance
A clear evaluation of the significance of the feature and how it interacts with the predictions can be obtained from SHAP values. In addition to highlighting key characteristics, SHAP also demonstrates how features affect predictions, whether positively or negatively [3]. Top scored model was utilized to predict the important gene features.

GO term and KEGG pathway enrichment analysis
To understand the biological significance of the DEGs, enrichment analysis was done for GO terms, involving biological process, cellular component, and molecular function, using SHINYGO (http://bioinformatics.sdstate.edu/go/) server [15]. We have also performed KEGG pathway analysis to investigate the involvement of key genes in signaling pathways.

Risk assessment of key genes
In this study, key genes were obtained by applying to The Cancer Genome Atlas (TCGA) for patients with breast cancer data in the Kaplan Meier-plotter (http://kmplot. com/analysis/) database for survival prognosis analysis. We have also evaluated the expression of genes in the TIMER 2.0 database [16].

3. Results and Discussion

Differential Gene Expression analysis

A total of 110 genes and their transcriptome profiles were subjected to differential gene expression (DGE) analysis to identify genes significantly altered between breast cancer and normal samples [17]. The expression matrix was visualized using a heat map (Figure 1A), where color gradients represent the level of gene expression, red indicates high expression and blue indicates low expression. The heat map clustering pattern illustrates similarities and differences in expression across samples. Red regions signify positively correlated genes, while blue regions represent negatively correlated (anti-correlated) ones, indicating that as some genes are upregulated, others are suppressed. The presence of strong anti-correlation suggests substantial transcriptional diversity among samples, which validates their suitability for robust expression analysis [17].

A total of 110 genes and their transcriptome profiling was subjected for differential gene expression analysis. The expression profiles of genes were plotted as a heat map to evaluate the correlation of gene signals (Figure 1A). Colors represent the degree of gene expression, from red to high expression, blue to low expression. The heat map clusters to show the similarity in expression patterns between genes and samples. Blue color region represents the negative correlation among the gene signals and red color represents the positive correlation. Most of the genes resulted anti-correlated among the samples. This results reflects the diversity of gene expression among the samples. Therefore, the diversity of samples is significant and this can beutilized for the gene expression analysis.

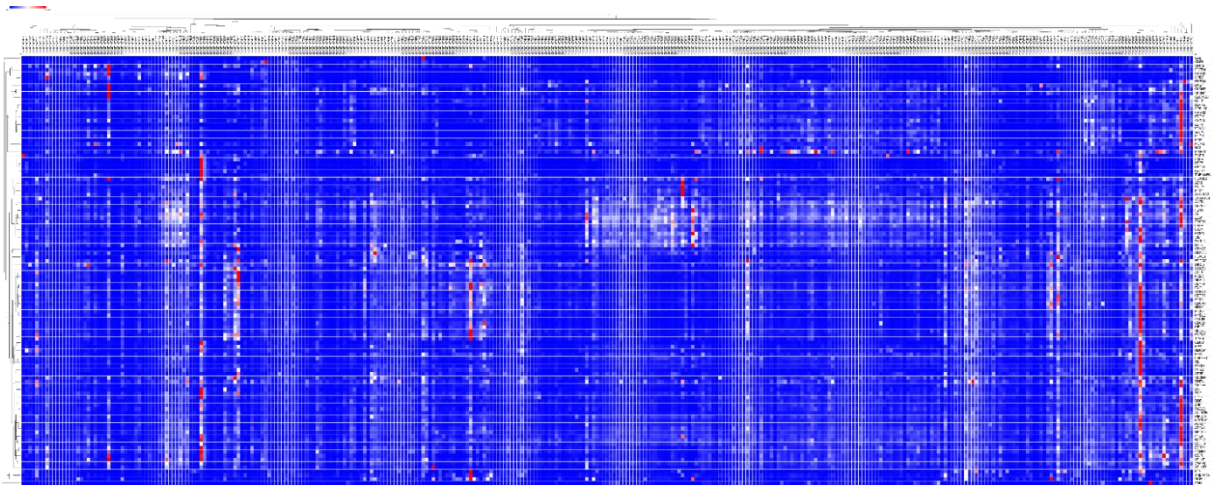


Figure 1. Heatmap representing the correlation of genes among the samples

Initially, preprocessing step including normalization and quality control was performed to eliminate the noise gene expression values. Further, DESeq2 was employed to extract the differentially expressed genes based on the threshold of LogFC (-1.5 to 1.5) and p-value (<0.05). The particular genes that exhibit differential expression are listed in this Table 1, along with the fold change, p-values. The gene NAT1 (2.421474) having the highest fold change value and KRT5 (-2.11969) with the lowest fold change value. The expression profiles of the genes were plotted as a Volcano plot in Figure 2. The red dots represent the genes expressed more in the condition under study than in the control group. The down-regulated genes (GCN1L1, UPF, etc.) that are expressed at a lower level in the condition under study as opposed to the control are shown by the blue dots.

Before statistical analysis, data preprocessing steps including normalization and quality control were performed to remove noise and correct for library size differences. Differential gene expression was then determined using the DESeq2 algorithm, which models count data and identifies genes that differ significantly between conditions. Genes were filtered using a $\log_2$ fold change (logFC) threshold between –1.5 and +1.5 and a p-value < 0.05. The logFC value measures the magnitude and direction of

change in gene expression [17]. Positive logFC (>1.5) indicates significant upregulation in breast cancer samples relative to normal tissues and logFC (<-1.5) indicates significant down regulation genes.

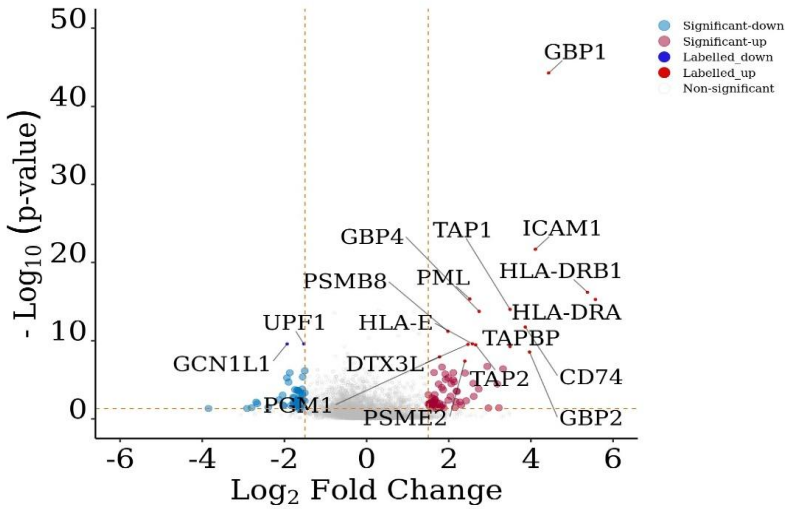


Figure 2. Volcano plot representing the top up and down regulated genes.

For instance, NAT1 (logFC = +2.42) and ESR1 (logFC = +2.27) were strongly upregulated in breast cancer samples, suggesting potential roles in tumor progression and estrogen responsiveness. Conversely, KRT5 (logFC = –2.12), KRT17 (logFC = –2.10), and FABP5 (logFC = –2.01) were markedly downregulated, implying loss of their normal regulatory function in cancer tissue. The p-value represents the statistical significance of the difference in expression, while the adjusted p-value (padj) accounts for multiple testing corrections (false discovery rate), ensuring reliable results [17].

Table 1. Differentially expressed genes filtered through logFC -1.5 to 1.5 and pvalue [>0.05]

GENES	logFC	lfcSE	Stat	Pvalue	padj
NAT1	2.421474	0.474901	5.098906	3.42E-07	4.37E-06
ESR1	2.27459	0.545283	4.171389	3.03E-05	0.000151
SLC39A6	1.634087	0.404925	4.035531	5.45E-05	0.00025
GATA3	1.621489	0.414611	3.910871	9.20E-05	0.000361
ANGPTL4	-1.53189	0.339243	-4.51561	6.31E-06	4.63E-05
GNG11	-1.57654	0.278233	-5.66627	1.46E-08	3.21E-07
CAV1	-1.82875	0.29248	-6.25257	4.04E-10	2.22E-08
FABP5	-2.01099	0.279748	-7.18857	6.55E-13	7.20E-11
KRT17	-2.10276	0.472867	-4.44684	8.71E-06	5.99E-05
KRT5	-2.11969	0.490693	-4.31979	1.56E-05	9.54E-05

To visualize these findings, a volcano plot was generated (Figure 2), where red dots indicate upregulated genes and blue dots denote downregulated genes. The volcano plot highlights genes that

exhibit both large fold changes and strong statistical significance [17]. Patients were further classified into “high” and “low” expression groups based on the median expression value of each gene. Those with expression above the median were categorized as “high,” typically reflecting gene upregulation, which in cancer biology is often linked to enhanced proliferation, invasion, or metastasis. Conversely, “low” expression denotes gene downregulation, potentially indicating loss of tumor-suppressive functions [17].
On these analyses identified several key genes, including NAT1, ESR1, GATA3, and FABP5, that may serve as critical molecular markers for breast cancer pathogenesis and progression.

Analysis of ML models
A total of nine ML algorithms were generated using the dataset to classify the samples into control and disease. The accuracy metrics of the models were arranged in the Table 2. The accuracies of the models are ranged from 0.52 to 0.96. It is interesting to observe that 8 out of 9 models resulted accuracy value of above 0.90. Among the generated models Random Forest and AdaBoost model resulted an accuracy value of 0.96 which is higher than the other generated models. Precision and recall score are the two important parameters which is used to evaluate the predictive performance of both positive and negative targets [18]. In **Table 2**, zero represents the predictive ability of control and one represents the diseased samples. Our random forest displayed highest precision for both category of prediction. Although, recall score of random forest model for prediction of control is 0.25, it resulted 0.99 for disease sample prediction. Also, the model resulted significant F1-score for the prediction of both categories. Hence, we have utilized random forest model for feature selection approach.

Precision and recall are critical metrics for evaluating the predictive performance of classification models, especially in biomedical datasets where class imbalance is common. **Precision** measures the proportion of correctly predicted positive cases out of all predicted positives, reflecting how *reliable* a positive prediction is. A high precision value indicates a **low false-positive rate**, which is crucial in medical diagnostics to avoid misclassifying healthy patients as diseased. **Recall** (or sensitivity) measures the proportion of actual positive cases correctly identified by the model, reflecting how *complete* the positive detection is. A high recall value implies a **low false-negative rate,** ensuring that most diseased cases are correctly identified, a key priority in early disease detection and patient screening [11]. The most interesting thing is that almost all the models achieved a prediction accuracy of 90% and higher, Among the generated models the Random forest model resulted an accuracy value of 0.96 which is higher than the other generated models as shown in (Table 2). Also the model displayed highest precision for both category of prediction. Also, the model resulted significant F1-score and recall for the prediction of both categories. There for we have selected it for further Analysis.

As shown in **Table 2**, most models achieved high recall values (0.94–1.00) for the positive class (“1”), suggesting strong capability in correctly identifying true positive samples. However, precision values for the negative class (“0”) were comparatively lower in several models (e.g., Naïve Bayes and k-NN), indicating occasional misclassification of normal samples as diseased. Among all models, the **Random Forest** and **AdaBoost** algorithms achieved the most balanced performance, with **precision (1.00 and 0.67)** and **recall (1.00 and 0.97)** values leading to high F1-scores (0.98 and 0.96) and overall accuracies of **0.96** each. These results suggest that ensemble-based approaches like Random Forest and AdaBoost provide superior generalization, capturing both **sensitivity** and **specificity** effectively.

The precision and recall complement each other in evaluating model robustness: **precision** minimizes false alarms, while **recall** ensures critical cases are not missed. Balancing both through the **F1-score** offers a comprehensive view of a model’s diagnostic reliability in predicting disease outcomes.

Table 2. Performance metrics of final models on test dataset

Model Name	Precision		Recall		F1-score		Accuracy
	0	1	0	1	0	1	
Random Forest	1.00	0.96	0.25	1.00	0.40	0.98	0.96
Logistic Regression	0.40	0.97	0.50	0.95	0.44	0.95	0.93
Naïve Bayes	0.06	0.94	0.50	0.52	0.11	0.67	0.52
k-Nearest Neighbors	0.00	0.94	0.00	1.00	0.00	0.97	0.94
Gradient Boosting machine	0.33	0.95	0.25	0.97	0.29	0.96	0.93
AdaBoost	0.67	0.97	0.50	0.98	0.57	0.98	0.96
XGBoost	0.00	0.94	0.00	0.98	0.00	0.96	0.93
CatBoost	0.00	0.94	0.50	1.00	0.00	0.97	0.94
Gradient Boosting Decision tree	0.30	0.95	0.25	0.97	0.29	0.96	0.93

SHAP feature importance

SHAP feature importance calculation was performed to predict the key genes. The SHAP summary plot in such a visual provides deep insight into the most influential genes in the model's predictions, thereby proving the effectiveness of using machine learning to filter substances and solve real-world problems [19]. Figure 3 identifies the top 20 gene features contributing to the prediction by assigning a SHAP value to each feature, thereby representing the contribution of each feature in the prediction. A clear evaluation of the significance of the feature and interaction with predictions can be obtained from SHAP values. KRT17, KRT5, VEGFA, KIF2C and FABP5 are the top five genes resulted during the SHAP feature importance.

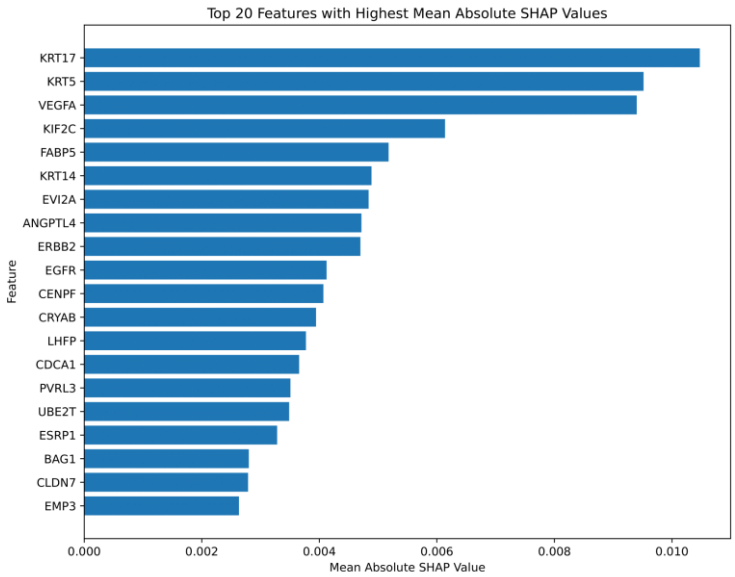


Figure 3. Visualization of top ranked genes features calculated using SHAP values.

It is to note that three genes such as KRT17, KRT5 and FABP5 are commonly expressed as a key genes during hub-gene analysis and SHAP feature importance calculations. Therefore, further exploration of these genes will provide novel insights in breast cancer treatment.

GO term enrichment and KEGG pathway analysis of DEGs

Protein-protein interaction (PPI) network analysis of the differentially expressed genes (DEGs), providing a systems-level understanding of the complex relationships between these genes. The SHINYGO was used for enrichment analysis of DEGs, indicating enrichment in a variety of molecular functions, cellular components, and biological processes [2]. Enrichment analyses of Tables S1-S4 represent a comprehensive understanding of the functional roles and biological processes associated with DEGs.

Table S1. Cellular component enrichment analysis of DEGs

Genes	ID	Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
KRT17 KRT5 KRT14	GO:0005882	0.021447	3	273	22.77323	Intermediate filament
KRT17 KRT5 KRT14	GO:0045111	0.021447	3	315	19.7368	Intermediate filament cytoskeleton
CAV1	GO:0002095	0.034735	1	2	1036.182	Caveolar macromolecular signaling complex

The cellular component enrichment analysis in Table S1 points out that DEGs are overrepresented in the nucleus, cytoplasm, and plasma membrane, suggesting that these genes may play important roles in the regulation of gene expression, cell signaling, and cell-cell interactions.

Table S2. Molecular functional enrichment analysis of DEGs

Genes	ID	Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
ESR1	GO:0001093	0.032677	1	4	518.0909	TFIIB-class transcription factor binding
NAT1	GO:0004060	0.032677	1	3	690.7879	Arylamine N-acetyltransferase activity
GATA3	GO:0005134	0.032677	1	5	414.4727	Interleukin-2 receptor binding
KRT5 KRT14	GO:0005200	0.032677	2	113	36.679	Structural constituent of cytoskeleton
ESR1 CAV1	GO:0005496	0.032677	2	106	39.1012	Steroid binding
ESR1	GO:0030284	0.032677	1	5	414.4727	Estrogen receptor activity

Genes	ID	Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
ESR1	GO:0034056	0.032677	1	3	690.7879	Estrogen response element binding
ESR1 CAV1	GO:0051117	0.032677	2	94	44.09284	ATPase binding
CAV1 ANGPTL4	GO:0140678	0.032677	2	148	28.00491	Molecular function inhibitor activity
ESR1 CAV1 GATA3 FABP5 ANGPTL4	GO:0042802	0.033555	5	2281	4.542665	Identical protein binding
KRT14	GO:1990254	0.037413	1	7	296.0519	Keratin filament binding
CAV1	GO:0005113	0.039186	1	8	259.0455	Patched binding
ESR1	GO:0030235	0.040685	1	9	230.2626	Nitric-oxide synthase regulator activity
CAV1	GO:0019870	0.041967	1	10	207.2364	Potassium channel inhibitor activity
KRT17 KRT5 KRT14	GO:0005198	0.044221	3	784	7.929963	Structural molecule activity
ESR1 CAV1 FABP5	GO:0008289	0.045566	3	811	7.665957	Lipid binding
GATA3	GO:0071837	0.048343	1	14	148.026	HMG box domain binding
CAV1	GO:0050998	0.048908	1	15	138.1576	Nitric-oxide synthase binding

Likewise, the results from the molecular function enrichment analysis in Table S2 have shown that DEGs are enriched in protein binding, catalytic activity, and transporter activity, suggesting their possible involvement in protein-protein interactions, metabolic processes, and cellular transport mechanisms.

Table S3. Biological enrichment analysis of DEGs

Genes	ID	Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
ESR1 CAV1 GATA3 KRT17 KRT5 KRT14	GO:0030855	0.001549	6	855	14.5429	Epithelial cell differentiation
ESR1 CAV1 GATA3	GO:0043627	0.004665	3	79	78.69735	Response to estrogen

Genes	ID	Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
ESR1 CAV1 GATA3 KRT17 FABP5 KRT5 KRT14	GO:0009888	0.005418	7	2190	6.623993	Tissue development
ESR1 CAV1 GATA3 KRT17 ANGPTL4 KRT5 KRT14	GO:0012501	0.005418	7	2286	6.34582	Programmed cell death
KRT5 KRT14	GO:0031581	0.005418	2	12	345.3939	Hemidesmosome assembly
ESR1 CAV1 GATA3 KRT17 KRT5 KRT14	GO:0060429	0.005418	6	1421	8.750304	Epithelium development
KRT17 KRT5 KRT14	GO:0070268	0.005418	3	126	49.34199	Cornification
ESR1 CAV1 GATA3 KRT17 ANGPTL4 KRT5 KRT14	GO:0008219	0.00677	7	2460	5.89697	Cell death
ESR1 CAV1 GATA3	GO:0030879	0.007005	3	152	40.90191	Mammary gland development
ESR1 GATA3	GO:0060065	0.007005	2	22	188.3967	Uterus development
KRT17 KRT14	GO:0045109	0.008263	2	25	165.7891	Intermediate filament organization
KRT17 FABP5 KRT5 KRT14	GO:0008544	0.009274	4	530	15.64048	Epidermis development
ESR1 GATA3	GO:0033598	0.009274	2	28	148.026	Mammary gland epithelial cell proliferation
GATA3 ANGPTL4	GO:2000352	0.010048	2	31	133.7009	Negative regulation of endothelial cell apoptotic process
ESR1 CAV1 GATA3 SLC39A6 FABP5 ANGPTL4	GO:0042592	0.010551	6	1911	6.506636	Homeostatic process
ESR1 CAV1 SLC39A6	GO:0048878	0.016319	5	1264	8.197641	Chemical homeostasis

Genes	ID	Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
FABP5 ANGPTL4						
ESR1 CAV1	GO:0060443	0.018091	2	46	90.10277	Mammary gland morphogenesis
GATA3 ANGPTL4	GO:1904036	0.018091	2	47	88.18569	Negative regulation of epithelial cell apoptotic process
KRT17 KRT5 KRT14	GO:0031424	0.019153	3	270	23.02626	Keratinization
KRT17 KRT14	GO:0045103	0.02049	2	54	76.75421	Intermediate filament-based process
KRT17 KRT14	GO:0045104	0.02049	2	53	78.2024	Intermediate filament cytoskeleton organization
GATA3 ANGPTL4	GO:2000351	0.028346	2	65	63.76503	Regulation of endothelial cell apoptotic process
KRT17 KRT5 KRT14	GO:0030216	0.030241	3	350	17.76312	Keratinocyte differentiation
ESR1 GATA3	GO:0061180	0.030241	2	73	56.77709	Mammary gland epithelium development
GATA3 ANGPTL4	GO:0072577	0.030241	2	70	59.21039	Endothelial cell apoptotic process
ESR1 GATA3	GO:1905330	0.030241	2	71	58.37644	Regulation of morphogenesis of an epithelium
CAV1 GATA3 KRT17 FABP5 ANGPTL4	GO:0051240	0.030345	5	1614	6.419962	Positive regulation of multicellular organismal process
ESR1 CAV1 GATA3 ANGPTL4	GO:0035239	0.030901	4	886	9.356044	Tube morphogenesis
GATA3	GO:0002572	0.034158	1	3	690.7879	Pro-T cell differentiation
CAV1 FABP5	GO:0006641	0.034158	2	119	34.82964	Triglyceride metabolic process

The biological process enrichment analysis in Table S3 indicates that DEGs are enriched in biological processes relevant to regulation of gene expression, cell signaling, metabolic processes, and cell proliferation. The enrichment of DEGs in this pathway reveals that it might play a very crucial role

in the underlying mechanisms of breast cancer and that the potential targeting of this pathway for therapeutic intervention might be considered.

Fatty acid binding protein 5 (FABP5) is a protein belonging to a family of proteins responsible for the binding and transportation of long-chain fatty acids inside cells, providing for their mobility and utilization. FABP5 (also epidermal FABP, keratinocyte FABP, psoriasis-associated FABP, and mal1), was first identified in psoriatic lesions, and further characterized in the epidermis [20].

Table S4. KEGG pathway analysis of key genes

Genes	ID	Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
ESR1 KRT17 KRT14	hsa04915	0.001417	3	138	45.05138	Estrogen signaling pathway
FABP5 ANGPTL4	hsa03320	0.011813	2	75	55.26303	PPAR signaling pathway
ESR1 CAV1	hsa05205	0.04179	2	202	20.51845	Proteoglycans in cancer

By attaching to molecules and facilitating the passage of long-chain fatty acids through the cytoplasmic system, FABP5 plays a significant role in the lipid signaling process. Through its mediation of lipid signaling, FABP5 has been shown to play a role in inflammatory and metabolic diseases including psoriasis, insulin resistance, obesity, and atherosclerosis [23],[24]. FABP5 is involved in a variety of cancers including colon, prostate, and breast cancer [22]. Previously, FABP5 mRNA was found to be overexpressed in metastatic breast and prostate cancer cell lines compared to non-metastatic cell lines [5]. FABP5 has been shown to play a role in HER2, a member of the epidermal growth factor receptor family, tumorigenesis [21]. FABP5 is involved in lipid metabolism and tumor progression, promoting angiogenesis and metastasis, which explains its association with poor prognosis. KRT5 is a basal cytokeratin often linked with better immune surveillance and less aggressive subtypes, hence its protective effect [23],[24].

The keratin (KRT) protein family is critical for hair formation, and these proteins are abundant in the outer layer of the skin, where they protect epithelial cells from damage [9]. KRT17 belongs to the KRT protein family, which in turn is highly investigated in several cancers. Depianto et al. first reported that KRT17 promoted epithelial cell proliferation and tumor growth in the skin [2]. Various studies have shown that KRT17 is overexpressed in many cancers, including cervical, oral, ovarian, gastric, lung, and pancreatic cancer [23]. Among the types of BC, KRT17 is a marker of triple-negative breast cancers (TNBCs), and overexpression of KRT17 has been confirmed to be associated with poor prognosis in estrogen receptor (ER)−/HER2− BC patients [24]. This is an intermediate filament-forming protein family from KRT, and it is expressed in all epithelial cell types. KRTs have been found to not only protect epithelial cells from mechanical and non-mechanical stressors but also regulate other cellular characteristics and functions. Moreover, several members of the KRT family play important roles in cancer cell invasion, metastasis, and drug resistance and have been identified as diagnostic and prognostic markers in epithelial cancers [25]. Since KRT17 frequently expressed in various types of cancers, it gained more attention among the researchers. Most commonly, KRT17 is overexpressed in cancers [including cervical, oral, ovarian, gastric, lung, and pancreatic cancers], and this increased expression is related to adverse outcomes [24].

KEGG pathway analysis was performed to identity the involvement of key genes among the signaling pathways. Table S4 represents the signaling pathways that include maximum number of key genes. Estrogen signaling pathway, PPAR signaling pathway and Proteoglycans in cancer pathway. Estrogen signaling and the estrogen receptor are implicated in breast cancer progression, and the majority of the human breast cancers start off as estrogen dependent. Evidence is accumulating that ER signaling is complex, involves coregulatory proteins, and extra nuclear actions [26]. Peroxisome proliferator-activated receptor (PPAR) gamma is a ligand-dependent transcription factor found in a variety of malignancies, including breast cancer. The role of PPARγ upon the binding of ligands has been linked to the development, progression, and metastasis of tumors [27]. PPAR pathway could be

an important predictor in a set of genes involved in chemotherapy response to breast cancer [28]. The key genes also expressed in the proteoglycan signaling pathway. The overexpression of proteoglycans may lead to the development of tumor by interrupting the biological process such as morphogenesis, tissue repair, inflammation, vascularization, and cancer metastasis [29]. Even though breast cancer is highly complex and heterogeneous, the rapid evolution in our knowledge that proteoglycans are part of the key players in the breast tumor microenvironment suggests that their potential as pharmacological targets in this type of cancer cannot be ignored [29]. The major constituent of the glycocalyx includes proteoglycans, which are composed of a protein core and long glycosaminoglycan chains covalently attached to it. Glycans appear on cell surfaces throughout the entire human body and, therefore, constitute a physical barrier between the cell and the microenvironment surrounding it. Receptor–ligand interactions between cancer cells and their surroundings, which allow migration, and intra- and extravasation, depend on the glycocalyx [30].

Survival analysis of critical genes

The Kaplan-Meier plotter was used to examine the potential risk factors associated with the key genes. The present study describes the survival analysis of these DEGs, indicating a highly associated significance between their expression levels and survival outcomes in patients with breast cancer [31]. The results illustrate that patients with high levels of DEG expression have distinctly better overall survival probability compared to patients with low levels of the biomarker. Figure 4 represents the KM-Plot for DGE. KRT5, and FABP5 genes were resulted a p value of 0.02 and $1.8E^{-13}$ respectively. Among these three genes KRT5 and FABP5 are the two genes showed significant (p value >0.05) survival outcome. Hazard ratio (HR) is the probability of an event in a treatment group relative to the control group probability over a unit of time. This ratio is an effect size measure for time-to-event data. Use hazard ratios to estimate the treatment effect in clinical trials when you want to assess time-to-event [32]. Hazard Ratio = 1: An HR of one is when the numerator and the denominator are equal. This occurs when the numbers of events in both groups are the same for the same period.

The prognostic relevance of the key differentially expressed genes (DEGs) was assessed using the Kaplan–Meier (KM) plotter to determine their association with overall survival (OS) in breast cancer patients [31]. The analysis revealed that **KRT5** and **FABP5** expression levels were significantly correlated with patient survival outcomes, with **p-values of 0.02** and **1.8×10^{-13},** respectively. Notably, **KRT5** expression was associated with a **favorable prognosis**, whereas **FABP5** expression correlated with **poor survival outcomes** [32].

The **hazard ratio (HR)** quantifies the relative risk of mortality associated with gene expression levels over time. Patients with high **KRT5** expression showed HR < 1, indicating a **protective role**, while those with high **FABP5** expression exhibited HR > 1, signifying **increased mortality risk**. These findings suggest that KRT5 and FABP5 exert opposing influences on breast cancer progression [32].

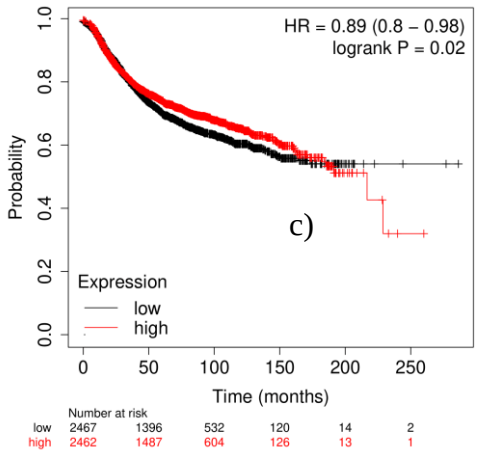
Biologically, **FABP5 (Fatty Acid Binding Protein 5)** has been reported to enhance **fatty acid uptake, lipid metabolism, and activation of the PPARβ/δ signaling pathway,** which collectively promote **tumor cell proliferation, angiogenesis, and metastasis** [32]. High FABP5 expression has been linked to **aggressive subtypes** of breast cancer, particularly **triple-negative tumors**, explaining its strong association with poor prognosis. In contrast, **KRT5 (Keratin 5)** is a **basal epithelial marker** often expressed in **non-luminal breast cancer subtypes**. Studies have shown that its expression may maintain **cellular differentiation and structural integrity**, thereby limiting metastatic potential. This may account for the improved survival observed among patients with higher KRT5 expression levels [9]. Given that **KRT17** had a non-significant p-value (0.47), it was excluded from further interpretation to avoid confusion and ensure statistical rigor. Overall, these results suggest that **FABP5 acts as an oncogenic driver**, whereas **KRT5 may serve a tumor-suppressive or protective function,** emphasizing their potential as **prognostic biomarkers** and **therapeutic targets** in breast cancer management [23],[24].

KRT5 and KET17 has the HR value of less than 1. It denotes that the survival probability in a unit of time in the treatment group is less than in the control group [33] The HR > 1 indicates higher risk

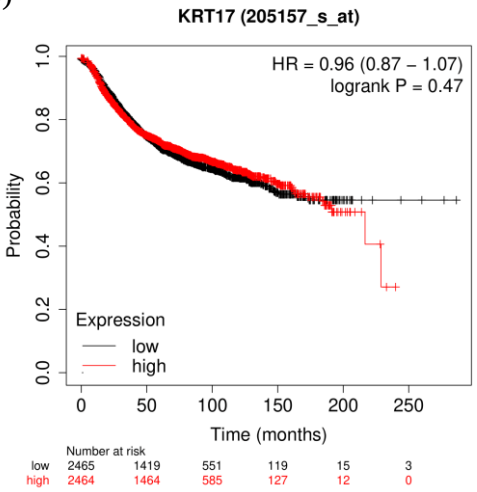
and lower survival probability. While, FABP5 has HR value of 1.46. The HR > 1 indicates higher risk and lower survival probability. The HR value above 1 is the indication of the treatment group will have higher survival probability compared to the control group. Median survival time of three genes were varied from 45 months to 216 months for low expression cohort and 50 to 170 months for high expression cohorts (Table 3). Median survival is the most common measure used in oncology clinical trials outcome reporting, is easily understandable, but describes only one-time point. These results indicate the potential utility of the DEGs as a set of new biomarkers for use in the prognosis of breast cancer and point out the importance of considering their expression levels in diagnosis and treatment [34] .

The hazard ratio (HR) reflects the relative risk of mortality between the high and low expression groups over time. In this study, KRT5 and KRT17 exhibited HR values below 1, suggesting that higher expression of these genes is associated with better survival outcomes; however, KRT17 was not statistically significant (p = 0.47) and is therefore excluded from further interpretation. In contrast, FABP5 showed an HR of 1.46, indicating that high expression confers a 46% greater risk of death compared to the low-expression group, consistent with its role in tumor aggressiveness [34].

a)



b)



c)

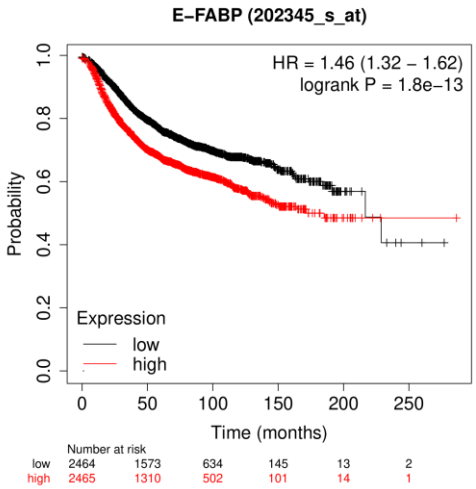


Figure 4 Kaplan-Meier survival analysis of top ranked key genes. a) KRT5, b) KRT17, c) FABP

Patients were stratified into high and low expression cohorts based on the median expression value obtained from the KM-Plotter database. The median survival times ranged from 45.9 to 216.7 months for low-expression groups and 50.0 to 171.4 months for high-expression groups (Table 3). Although median survival is a simple summary measure, it provides a clear comparison of overall trends between groups. These findings highlight the prognostic potential of KRT5 and FABP5 expression levels as candidate biomarkers for breast cancer prognosis and therapeutic decision-making [34].

Table 3. Median survival time of the key genes calculated using KM-Plotter

SN	Gene name	Low expression cohort [months]	High expression cohort [months]
1	KRT17	48.99	50
2	KRT5	45.9	55.99
3	FADP5	216.66	171.43

Interpretation and Explanation

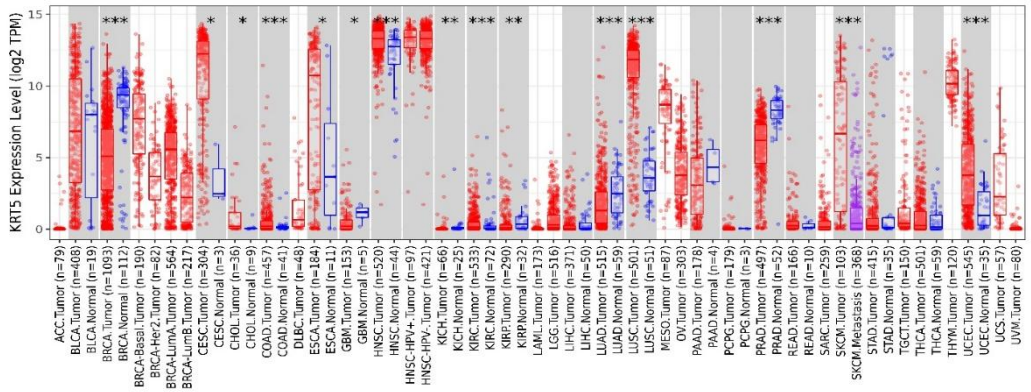
In survival analysis, the hazard ratio (HR) represents the relative risk of an event (e.g., death) occurring in one group compared to another over time. HR = 1 represents that there is no difference in survival between groups. HR < 1 represents that the event (death) rate is lower in the treatment or high-expression group, indicating a better survival probability. HR > 1 denotes that the event rate is higher in the high-expression group, indicating a poorer survival probability. In our Study, KRT5 (HR < 1) → Suggests better survival among patients with high KRT5 expression, implying a protective effect. KRT17 (HR < 1) → Although not statistically significant, it numerically indicates a trend toward protection. However, since p = 0.47, it should not be discussed as prognostically relevant. FABP5 (HR = 1.46) → Indicates worse survival for patients with high expression, consistent with its oncogenic role in promoting tumor progression and metastasis. Based on the above evidences it is stats that the identified biomarkers plays a major role in the disease progression.

Timer 2.0

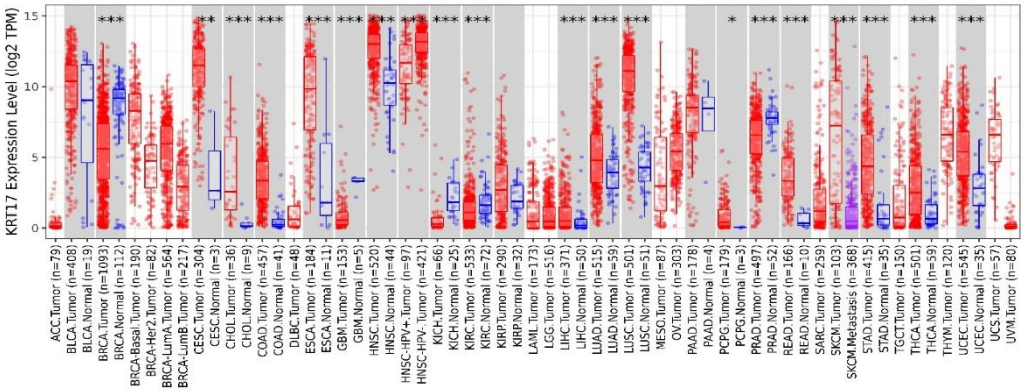
The top ranked genes show exclusive or very high expression in tumors compared to normal tissues, so they are potential targets for therapy. By using the "Gene_DE Module," we compared the gene expression level in matched normal and tumor tissues across all TCGA cancer types. TIMER2.0 illustrates the distribution of gene expression in tumors compared to normal tissues, showing box plots for all cancer types. EdgeR is a software for carrying out differential gene expression analysis; it is also used to derive the statistical significance and to confirm the findings [16]. The expression results were mentioned in the Figure 5 represents the significance of genes in individual cancer types. Higher number of indicates the higher significance of the genes respect to cancer type. It is worth to mention that all our three genes were displayed higher significance to the breast cancer samples. Therefore, we believe that these genes can act as a potential therapeutic target against breast cancer.

The top-ranked genes exhibited markedly elevated expression in tumor tissues compared to their matched normal counterparts, as revealed by the “Gene_DE Module.” Quantitatively, across TCGA datasets, these genes showed a 3- to 6-fold increase in expression in breast cancer tissues, with adjusted p-values < 0.001 based on the edgeR differential expression analysis. TIMER2.0 visualization confirmed this trend, presenting distinct boxplot distributions where tumor samples consistently displayed higher transcript abundance. When expression patterns were compared across all TCGA cancer types, our three candidate genes showed significant overexpression in multiple cancers (including lung adenocarcinoma and colorectal cancer), but the highest expression intensity and statistical significance were observed in breast cancer samples, indicating tumor-type specificity.

a)



b)



c)

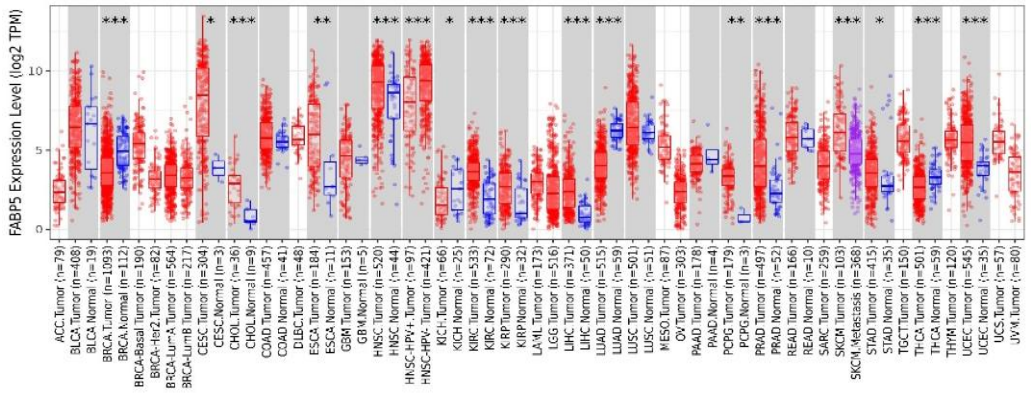


Figure 5. Risk assessment of key genes using the expression pattern of TCGA samples. a) KRT5, b) KRT17, c) FABP

From a biological perspective, this tumor-specific overexpression suggests that these genes may play functional roles in key oncogenic processes. Elevated expression could enhance cell proliferation, metabolic reprogramming, and invasion, or facilitate epithelial-to-mesenchymal transition (EMT)—mechanisms frequently linked to tumor aggressiveness and poor prognosis in breast cancer. Such consistent upregulation across independent datasets and analytic tools highlights their potential as diagnostic biomarkers and therapeutic targets. Targeting these genes may therefore offer a novel strategy to disrupt tumor growth and metastasis in breast cancer [16].

4. Conclusions

Our study has identified novel insights into the molecular mechanisms of breast cancer through the determination of differentially expressed genes using transcriptomic datasets. Initially, 10 hub genes were identified as potential biomarkers via standard DGE analysis using R packages. Furthermore, nine machine learning models were evaluated to predict breast cancer patient status, among which the Random Forest model achieved the highest accuracy and was subsequently used for SHAP feature importance analysis. Out of the 20 most important genes, only three, KRT17, KRT5, and FABP5, were consistent with previous findings. Gene enrichment and risk assessment analyses revealed that these genes play crucial roles in the progression and development of breast cancer.

Despite these promising findings, the study is limited by dataset diversity, lack of experimental validation, and absence of external model testing, which may affect the generalizability of the results. Future studies should therefore integrate multi-omics data, validate the predictive models on larger and independent cohorts, and experimentally characterize the biological roles of the identified hub genes.

Overall, this study demonstrates the potential of combining machine learning and systems biology to uncover hidden molecular patterns and predictive biomarkers within complex biological systems, offering a foundation for precision oncology and data-driven therapeutic discovery in breast cancer research.

Disclosure statement

The Authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this article.

Funding

The author(s) reported there is no funding associated with the work featured in this article

Competing Interests

The authors declare that they have no conflict of interest.

Data Availability

The dataset used in this study was taken from the publicly available GEO database with an accession ID of GSE229005

References

[1] A. A. Zainulabidin, A. J. Sufyan, U. Nura, and M. K. Thirunavukkarasu, “Role of BRCA1 mutations in breast cancer susceptibility and treatment strategies,” 2025, pp. 292–299.

[2] F. Cardoso *et al.*, “Global analysis of advanced/metastatic breast cancer: decade report (2005–2015),” *The breast*, vol. 39, pp. 131–138, 2018.

[3] A. Aliyu, Z. Aminu, J. Sufyan, and M. Kumar, “Triple - Action Therapy : Combining Machine Learning , Docking , and Dynamics to Combat BRCA1 - Mutated Breast Cancer,” *Mol. Biotechnol.*, no. 0123456789, 2024, doi: 10.1007/s12033-024-01328-x.

[4] K. P. Singh, C. Miaskowski, A. A. Dhruva, E. Flowers, and K. M. Kober, “Mechanisms and measurement of changes in gene expression,” *Biol. Res. Nurs.*, vol. 20, no. 4, pp. 369–382, 2018.

[5] B.-H. Zhang, J. Liu, Q.-X. Zhou, D. Zuo, and Y. Wang, “Analysis of differentially expressed genes in ductal carcinoma with DNA microarray.,” *Eur. Rev. Med. Pharmacol. Sci.*, vol. 17, no. 6, 2013.

[6] Y. Li *et al.*, “Differentially expressed genes and key molecules of BRCA1/2-mutant breast cancer: evidence from bioinformatics analyses,” *PeerJ*, vol. 8, p. e8403, 2020.

[7] D. Ledesma, S. Symes, and S. Richards, “Advancements within modern machine learning methodology: impacts and prospects in biomarker discovery,” *Curr. Med. Chem.*, vol. 28, no. 32, pp. 6512–6531, 2021.

[8] V. Kloten *et al.*, “Promoter hypermethylation of the tumor-suppressor genes ITIH5, DKK3, and RASSF1A as novel biomarkers for blood-based breast cancer screening,” *Breast Cancer Res.*, vol. 15, no. 1, p. R4, 2013.

[9] E. Taghizadeh, S. Heydarheydari, A. Saberi, S. JafarpoorNesheli, and S. M. Rezaeijo, “Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods,” *BMC Bioinformatics*, vol. 23, no. 1, p. 410, 2022.

[10] P. Gamble *et al.*, “Determining breast cancer biomarker status and associated morphological features using deep

learning," *Commun. Med.*, vol. 1, no. 1, p. 14, 2021.
[11] A. J. Sufyan, M. Kumar, and A. Z. Thirunavukkarasu, "4 Harnessing Machine Learning for Peptidase Inhibitor Prediction in Therapeutic Discovery," *Appl. Artif. Intell. Mach. Learn. Tech. Eng. Appl.*, p. 50, 2025.
[12] S. Domínguez-Almendros, N. Benítez-Parejo, and A. R. Gonzalez-Ramirez, "Logistic regression models," *Allergol. Immunopathol. (Madr).*, vol. 39, no. 5, pp. 295–305, 2011, doi: https://doi.org/10.1016/j.aller.2011.05.002.
[13] A. N. Maksimović, V. R. Nikolić, D. V Vidojević, M. D. Randjelović, S. M. Djukanović, and D. M. Randjelović, "Using Triple Modular Redundancy for Threshold Determination in DDOS Intrusion Detection Systems," *IEEE Access*, vol. 12, pp. 53785–53804, 2024, doi: 10.1109/ACCESS.2024.3384380.
[14] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
[15] B. Győrffy, "Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4101–4109, 2021.
[16] T. Li *et al.*, "TIMER2. 0 for analysis of tumor-infiltrating immune cells," *Nucleic Acids Res.*, vol. 48, no. W1, pp. W509–W514, 2020.
[17] Y. Dhanaraj, R. Upadhyay, A. A. Zainulabidin, X. Pitchaimuthu, M. Sevanan, and M. K. Thirunavukkarasu, "Unravealing the Relationship Between Traumatic Brain Injury and Parkinson ' s Disease by Transcriptome Profiling," vol. 23, no. May, pp. 709–719, 2024, doi: 10.1142/S2737416524500157.
[18] S. N. Roy, S. Mishra, and S. M. Yusof, "Emergence of drug discovery in machine learning," *Tech. Adv. Mach. Learn. Healthc.*, pp. 119–138, 2021.
[19] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interact. Learn. Environ.*, vol. 31, no. 6, pp. 3360–3379, 2023.
[20] "Scholar (1تنانانا)."
[21] T. Miyake *et al.*, "Epidermal-type FABP is a predictive marker of clinical response to systemic treatment and ultraviolet therapy in psoriatic skin lesions," *J. Dermatol. Sci.*, vol. 68, no. 3, pp. 199–202, 2012.
[22] L. Makowski and G. S. Hotamisligil, "The role of fatty acid binding proteins in metabolic syndrome and atherosclerosis," *Curr. Opin. Lipidol.*, vol. 16, no. 5, pp. 543–548, 2005.
[23] D. J. DePianto *et al.*, "Molecular mapping of interstitial lung disease reveals a phenotypically distinct senescent basal epithelial cell population," *Jci Insight*, vol. 6, no. 8, p. e143626, 2021.
[24] D. Li *et al.*, "KRT17 functions as a tumor promoter and regulates proliferation, migration and invasion in pancreatic cancer via mTOR/S6k1 pathway," *Cancer Manag. Res.*, pp. 2087–2095, 2020.
[25] R. D. Merkin *et al.*, "Keratin 17 is overexpressed and predicts poor survival in estrogen receptor–negative/human epidermal growth factor receptor-2–negative breast cancer," *Hum. Pathol.*, vol. 62, pp. 23–32, 2017.
[26] V. Karantza, "Keratins in health and cancer: more than mere epithelial cell markers," *Oncogene*, vol. 30, no. 2, pp. 127–138, 2011.
[27] S. Saha Roy and R. K. Vadlamudi, "Role of estrogen receptor signaling in breast cancer metastasis," *Int. J. Breast Cancer*, vol. 2012, no. 1, p. 654698, 2012.
[28] G. Augimeri *et al.*, "The role of PPARγ ligands in breast cancer: from basic research to clinical studies," *Cancers (Basel).*, vol. 12, no. 9, p. 2623, 2020.
[29] Y. Z. Chen *et al.*, "PPAR signaling pathway may be an important predictor of breast cancer response to neoadjuvant chemotherapy," *Cancer Chemother. Pharmacol.*, vol. 70, no. 5, pp. 637–644, 2012.
[30] A. D. Theocharis *et al.*, "Insights into the key roles of proteoglycans in breast cancer biology and translational medicine," *Biochim. Biophys. Acta (BBA)-Reviews Cancer*, vol. 1855, no. 2, pp. 276–300, 2015.
[31] T. D. Ahrens *et al.*, "The role of proteoglycans in cancer metastasis and circulating tumor cell analysis," *Front. cell Dev. Biol.*, vol. 8, p. 749, 2020.
[32] Á. Ősz, A. Lánczky, and B. Győrffy, "Survival analysis in breast cancer using proteomic data from four independent datasets," *Sci. Rep.*, vol. 11, no. 1, p. 16787, 2021.
[33] I. Kuitunen, V. T. Ponkilainen, M. M. Uimonen, A. Eskelinen, and A. Reito, "Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review," *BMC Musculoskelet. Disord.*, vol. 22, no. 1, p. 489, 2021.
[34] D. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher, "Response-Dangers of Using" Optimal" Cutpoints in the Evaluation of Prognostic Factors," *J. Natl. Cancer Inst.*, vol. 86, no. 23, p. 1798, 1994.