INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Unveiling Key Gene Biomarkers and Clinical Factors in Breast Cancer Subtypes Using Machine Learning and Cox Regression

Gitanjali Pradhani[1]*, **Jonali Gogoi**[2]

1 Research Scholar, Department of Statistics, Assam University, Silchar, Assam, India
2 Assistant Professor, Department of Statistics, Assam University, Silchar, Assam, India

OPEN ACCESS

Received: 29-04-2024
Accepted: 27-07-2024
Published: 14-08-2024

Citation: Pradhani G, Gogoi J (2024) Unveiling Key Gene Biomarkers and Clinical Factors in Breast Cancer Subtypes Using Machine Learning and Cox Regression. Indian Journal of Science and Technology 17(32): 3350-3360. https://doi.org/10.17485/IJST/v17i32.1440

* **Corresponding author**.

pradhanig18@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: Breast cancer is the most frequent malignancy among women, and understanding its genetic subtypes is crucial for improving patient outcomes. Using clinical and gene expression data, this study attempts to uncover important gene biomarkers and clinical factors for various subtypes of breast cancer. **Methods:** Ten machine learning techniques were initially employed to classify breast cancer subtypes, with the most accurate method serving as the base classifier. Eight feature selection techniques were then used to pinpoint important genes, which were combined with clinical characteristics and subjected to Cox regression analysis to determine their impact on overall survival. Optimal cut-points were identified for each significant gene to stratify data into high and low-expression groups, followed by Kaplan-Meier survival analysis. The selected genes underwent biological validation through Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, Gene Ontology (GO) enrichment analysis, and a protein-protein interaction (PPI) network. **Findings:** Age was found to be significantly associated with the hazard as well as survival of all breast cancer subtypes. The genes FOXP1, BRCA1, SETDB1, CHEK2, BRCA1, and MAP2K4 were most strongly associated with the survival time of Basal-Like, HER2, Claudin-Low, Luminal A, Luminal B, and Normal-Like subtypes, respectively. The PPI network identified CDKN2A as the hub protein with the highest node degree. KEGG pathway analysis revealed significant pathways, including the PI3K-Akt, FoxO and ErbB signalling pathways that are closely related to breast cancer. GO analysis identified numerous breast cancer-related biological processes, molecular functions, and cellular components. **Novelty:** This paper proposes a novel way to find important gene biomarkers and clinical factors by combining advanced feature selection methods with Cox regression. The results' robustness is further increased by applying KEGG Pathway and Gene Ontology analysis for biological validation.

Keywords: Breast cancer; Gene expression; Feature Selection; Machine Learning; Survival Analysis

1 Introduction

Studies on Cancer have emerged as a significant area in medical science over the past few decades [1]. It is both a genetic and environmental disorder that arises from a complex interplay between inherited genetic mutations and external environmental factors [2]. As a result, the effects of genes that regulate normal cellular processes—like growth, survival, and invasion/motility—are amplified. In contrast, the effects of genes that restrict these effects are repressed [3]. The rate of growth varies from cell to cell. Some cells grow faster while some have slower growth. However, any abnormal genetic mutation can disrupt apoptosis, also known as programmed cell death and may create an imbalance between cell division and cell death [4]. It thereby causes the tumor to develop in an uncontrolled manner. Hormone imbalances and hereditary genetic flaws are also important contributors to cancer growth. Ten most prevalent cancer types are responsible for over 70% of all cancer-related fatalities. Breast cancer in women is the most frequent cancer globally, accounting for 11.7% of all new cases of cancer. It is closely followed by lung cancer (11.4%), colorectal cancer (10.0%), prostate cancer (7.3%) and stomach cancer (5.6%) [5]. However, breast cancer has a good molecular classification system and targeted treatment options. The pure anatomical and pathological classification of breast cancer has been shifted to a new and fine classification based on molecular criteria due to the advancement in molecular biology.

Based on the molecular classifications, breast cancer can be mainly divided into six subgroups: normal-like, luminal A and B, HER2-positive, basal-like, and claudin-low. Basal-like and claudin-low subtypes, characterized by the lack of estrogen receptor (ER), progesterone receptor (PR), and HER2 expression [6].

The survival rate of breast cancer patients can be considerably increased by comprehensive treatment based on molecular typing, which includes surgery, radiation, chemotherapy, endocrine therapy, and targeted therapy. But the prognosis for patients is worse than anticipated because of the complex heterogeneity of the same molecular subtype. Moreover, the environmental etiologies that lead to breast cancer are still a mystery [7]. Thus, a comprehensive study of the similarities and variations between the biology and molecular pathways of patients along with the environmental factors is necessary to select more effective treatment strategies for the various molecular subtypes of breast cancer. However, the study focuses solely on determining the genes that are aberrantly expressed during the development of breast cancer along with the clinical factors and attempts to understand the molecular pathways, biological processes, and potential causes of the disease.

To the best of our knowledge, however, the subtype-specific classification of breast cancer is restricted in the relevant public datasets—aside from The Cancer Genome Atlas (TCGA). This fact drives our research into multiclass classification and inspires us to provide a strategy for choosing biomarkers for subtypes of breast cancer.

Recently, several studies have been utilizing machine learning approaches, including regression, classification, and clustering in order to try and identify the significant miRNAs and genes. Many studies have used validation techniques and gene expression analysis to identify important genes influencing breast cancer survival. Research has indicated that certain genes, such as COL11A1, MMP11, and COL10A1, are up-regulated in breast cancer, whereas other genes, such as PCOLCE2, LAMA2, TMTC1, ADAMTS5, TIMP4, and RSPO3, are down-regulated [8]. Crucial genes like RBP4 and CRYAB have also been found to be possible biomarkers for the diagnosis and prognosis of breast cancer. Additionally, a predictive gene model based on mRNAs has been developed, encompassing genes such as CFB, MAL2, PSME2, and others, indicating a possible correlation between immune modulation and specific somatic gene alterations in patients with breast cancer [9]. Furthermore, a predictive model that accurately predicts the 5-year overall survival of patients with breast cancer has been created which is based on transcription factors (TFs) such as ZIC2, POU3F2, PAX7, and others [10]. Together, these findings advance our knowledge of the molecular processes affecting breast cancer survival and offer insightful information for both clinical and future research.

Nonetheless, the majority of the studies only provide a limited amount of data on multi-class classification and regression for all types of cancer. These findings have led us to suggest a novel Machine Learning Integrated Ensemble of Feature Selection Methods, which we will use in conjunction with semi-parametric Cox regression-based survival analysis to determine the most important genes for subtypes of breast cancer based on gene expression and clinical data.

Using Breast Invasive Carcinoma (BRCA) data, machine learning with traditional cox model was used to identify the top genes associated with the survival of patients diagnosed with breast invasive carcinoma. Furthermore, survival analysis using the non-parametric Kaplan-Meier technique, protein-protein interaction network, KEGG pathway, and GO enrichment analysis to determine the biological importance of the chosen genes was carried out. Three sub-processes make up the remaining part of the method after the data is prepared. These include (a) the selection of a suitable machine learning technique to calculate classification accuracy, (b) selecting initial features using eight different feature selection methods, and (c) finally identifying the gene biomarkers using Cox regression survival analysis. Lastly, the study seeks to identify the common genes, unique differential molecules and signaling pathways between the cancer tissues and the cancer-adjacent normal tissues from the various subtypes of breast cancer, screen these genes for potential clinical applications, and assess their utility.

2 Materials and Methods

2.1. Data Preparation

The study utilized the Breast Invasive Carcinoma (BRCA) data from the cBio Cancer Genomics Portal which can be assessed from https://www.cbioportal.org/. It is an open-access resource for interactive exploration of multidimensional cancer genomics data sets[11]. Six molecular subtypes of breast cancer: Basal-Like, Claudin-Low, Luminal A, Luminal B, HER2-Enriched (Her2), and Normal-Like have been considered in this study. The OncoKBTM (https://www.oncokb.org/cancer-genes) Cancer Gene List's gene expression values based on next-generation sequencing (NGS) and extensive clinical data namely age at diagnosis, type of surgery (Mastectomy/ Breast Conserving), menopausal status (Pre/Post), tumor size and different therapies namely chemotherapy (Yes/No), hormone therapy (Yes/No) and radiation therapy (Yes/No) are included in the dataset. There was no need for additional processing because the raw mRNA expression levels were already log2-transformed to stabilize variance and normalized to z-scores to the expression distribution of all samples using the Illumina HT-12 v3 microarray platform. However, to balance the data, the Synthetic Minority Over-sampling Technique (SMOTE) was used because the distribution of the six subtypes was unbalanced[12].

2.2. Feature Selection Based on Machine Learning and Survival Analysis

2.2.1. Selection of a suitable machine learning technique

Using the gene expression data, which included six breast cancer subtypes, the best machine-learning method for classifying the subtypes was determined. Finding a gene set that could reliably divide the data into six different classes was the goal. Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbours (KNN), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Discriminant Analysis (DA), Gradient Boosting (GB), XGBoost (XGB), and LightGBM (LGB) are the ten well-known machine learning algorithms that were assessed. Using 10-fold cross-validation, each technique was applied to the pre-processed dataset. The best-performing machine learning methodology was chosen for the next steps in the suggested procedure based on the accuracy values obtained.

2.2.2. Feature selection using eight methods

Eight different feature selection methods were used on the dataset: Mutual Information Maximization (MIM), Minimum Redundancy Maximum Relevance (mRMR), Conditional Mutual Information Maximization (CMIM), Joint Mutual Information (JMI), Double Input Symmetrical Relevance (DISR), Interaction Capping (ICAP), Conditional Informative Feature Extraction (CIFE), and Conditional Redundancy (CONDRED). Every technique was used repeatedly to choose different gene subsets, progressively growing in size (e.g., top 5, top 10, top 15 genes, etc.). Following that, these subsets were examined using the selected machine learning method, and 10-fold cross-validation was used to gauge the classification performance. The greatest classification accuracy attained was used to define the ideal subset of genes for each feature selection technique. In the end, the intersecting gene subsets from every technique were found and chosen for additional examination based on which method produced the best accuracies for each subset size.

2.2.3. Survival analysis to identify gene biomarkers

The final step of the proposed method employs Cox regression-based survival analysis. During this process, Cox coefficients and p-values are calculated to pinpoint significant genes. Genes exhibiting p-values below the threshold of 0.05 are deemed significant and selected from the previously identified subsets for the various breast cancer subtypes. Moreover, pertinent clinical factors of the patients such as age, type of surgery, menopausal status, tumor size and different therapies namely chemotherapy, hormone therapy and radiation therapy are included in the multivariate Cox regression to assess their influence. This approach yields a thorough list of significant genes and clinical factors, which are subsequently considered as potential biomarkers for the breast cancer subtypes.

3 Results and Discussions

3.1. Selection of important features

LightGBM outperforms the other nine approaches with an accuracy of 90%. And a standard deviation of 0.01. As so, this experiment validates the choice of LightGBM for the next stages in the suggested procedure. The classification accuracies of the methods are given in Table 1.

Table 1. 10-fold cross-validated classification accuracy using all genes

Model	Accuracy	Standard Deviation
Support Vector Machine (SVM)	90 %	0.02
Artificial Neural Network (ANN)	90 %	0.02
K-Nearest Neighbors (KNN)	77 %	0.02
Decision Tree (DT)	73 %	0.02
Random Forest (RF)	89 %	0.02
Naive Bayes (NB)	74 %	0.02
Discriminant Analysis (DA)	77 %	0.02
Gradient Boosting (GB)	86 %	0.02
XGBoost (XGB)	89 %	0.02
LightGBM (LGB)	90 %	0.01

Following this, Table 2 reports the accuracy details for several subsets of genes, ranging from 5 to 50 genes. It is clear from the accuracy figures that various feature selection techniques identify different gene sets. For example, MIM outperforms other feature sets with an accuracy of 90.71 $\pm$ 0.03% when using a collection of 45 features. In a similar vein, CONDRED, mRMR, CMIM, JMI, DISR, CIFE, and ICAP produce the best accuracies of 90.83 $\pm$ 0.03%, 90.82 $\pm$ 0.03%, 90.98 $\pm$ 0.03%, 90.97 $\pm$ 0.03%, 90.98 $\pm$ 0.20%, 90.99 $\pm$ 0.03%, and 90.96 $\pm$ 0.03%, respectively, with 50 genes. The genes that are common across all these subsets are considered.

Table 2. Classification accuracy values produced by LightGBM for eight different feature selection methods

No. of Features	Mutual Information Maximization		Minimum Redundancy Maximum Relevance		Conditional Mutual Information Maximization		Joint Mutual Information	
	Accuracy	Std. Dev	Accuracy	Std. Dev	Accuracy	Std. Dev	Accuracy	Std. Dev
5	0.7449	0.0387	0.7454	0.0389	0.7453	0.0389	0.7011	0.0305
10	0.8382	0.0364	0.8391	0.0394	0.8391	0.0394	0.8306	0.0420
15	0.8606	0.0345	0.8620	0.0370	0.8620	0.0370	0.8628	0.0397
20	0.8739	0.0346	0.8746	0.0363	0.8745	0.0363	0.8819	0.0349
25	0.8955	0.0330	0.8924	0.0361	0.8924	0.0361	0.8899	0.0333
30	0.8960	0.0359	0.8982	0.0390	0.8982	0.0390	0.8987	0.0337
35	0.9018	0.0344	0.9032	0.0363	0.9032	0.0363	0.9042	0.0347
40	0.9023	0.0367	0.9025	0.0385	0.9024	0.0385	0.9055	0.0355
45	0.9071	0.0386	0.9057	0.0360	0.9057	0.0360	0.9060	0.0389
50	0.9040	0.0410	0.9083	0.0376	0.9082	0.0376	0.9098	0.0357
No. of Features	**Double Input Symmetrical Relevance**		**Conditional Informative Feature Extraction**		**Interaction Capping**		**Conditional Redundancy**	
	Accuracy	Std. Dev	Accuracy	Std. Dev	Accuracy	Std. Dev	Accuracy	Std. Dev
5	0.7011	0.0305	0.7012	0.0305	0.7010	0.0305	0.7010	0.0305
10	0.8305	0.0420	0.8306	0.0420	0.8310	0.0420	0.8304	0.0420
15	0.8627	0.0397	0.8628	0.0397	0.8627	0.0397	0.8626	0.0397
20	0.8818	0.0349	0.8819	0.0349	0.8820	0.0349	0.8817	0.0349
25	0.8899	0.0333	0.8900	0.0333	0.8898	0.0333	0.8898	0.0333
30	0.8987	0.0337	0.8988	0.0337	0.8988	0.0337	0.8986	0.0337
35	0.9042	0.0347	0.9043	0.0347	0.9043	0.0347	0.9041	0.0347
40	0.9054	0.0355	0.9055	0.0355	0.9056	0.0355	0.9053	0.0355
45	0.9060	0.0389	0.9061	0.0389	0.9061	0.0389	0.9059	0.0389
50	0.9097	0.0357	0.9098	0.0357	0.9099	0.0357	0.9096	0.0357

Subsequently, a multivariate Cox regression analysis was carried out, integrating all of the chosen genes with pertinent clinical variables. For every molecular subtype of breast cancer, a subtype-based Cox regression was performed to assess the influence of various variables and genes on the subtypes. The following Table 3 (**a, b, c, d, e, f**) display the coefficients, hazard ratios and p-values of the significant clinical variables and genes. For example, for the HER2 subtype, BRCA1 gene has a significant impact with the hazard rate of the patients with a p-value of 0.000. The cox coefficient value of 0.86 suggests that higher expression of BRCA1 is associated with a decreased risk of death which validates the fact that it is a tumor suppressor gene[13]. Specifically, a hazard ratio of 0.42 means that the hazard is 58% lower for patients with higher BRCA1 expression compared to those with lower expression.

Table 3. (a, b, c, d, e, f) Coefficients, hazard ratios, and p-values of the significant clinical variables and genes

(a) Her2 Subtype			
Covariate	**Coefficient**	**HR**	**p value**
Age	0.04	1.04	0.000
Menopausal State (Pre)	1.02	2.76	0.000
Chemotherapy (Yes)	0.60	1.82	0.002
BRCA1	-0.86	0.42	0.000
FOXO1	-0.66	0.52	0.000
NOTCH1	0.82	2.28	0.000
MTAP	-0.57	0.56	0.002
MAP3K1	-0.37	0.69	0.004
PPP2R2A	0.43	1.54	0.011
EGFR	0.25	1.28	0.012
JAK1	-0.43	0.65	0.013
TBX3	0.32	1.37	0.013
CCND3	0.45	1.57	0.017
FOXP1	0.59	1.81	0.019
CBFB	0.50	1.65	0.026
SMAD4	0.57	1.77	0.029
FOXO3	-0.25	0.78	0.030
AKT1	-0.28	0.75	0.037
HRAS	-0.31	0.74	0.046
(b) Basal Like Subtype			
Covariate	**Coefficient**	**HR**	**p value**
Menopausal State (Pre)	1.24	3.45	0.000
Age	0.07	1.08	0.000
BRAF	0.92	2.51	0.004
NRAS	0.70	2.01	0.001
AKT1	0.56	1.75	0.000
PIK3R1	0.38	1.46	0.043
CDKN2A	0.26	1.30	0.015
ERBB2	0.26	1.29	0.000
ARID5B	-0.45	0.64	0.006
SMAD4	-0.47	0.63	0.026
FOXP1	-1.27	0.28	0.007
(c) Claudin-Low Subtype (continues…)			
Covariate	**Coefficient**	**HR**	**p value**
Age	0.08	1.08	0.000
Menopausal State (Pre)	0.97	2.65	0.020
Chemotherapy (Yes)	0.62	1.86	0.024
SETDB1	-1.80	0.16	0.000
ERBB3	0.54	1.71	0.002
PPP2R2A	-1.12	0.33	0.007
MAP3K13	-0.88	0.42	0.017
CDH1	-0.27	0.76	0.026
CDKN2A	-0.40	0.67	0.027
CDKN1B	-0.63	0.53	0.028
(d) Luminal A Subtype			

Continued on next page

Table 3 continued

Covariate	Coefficient	HR	p value
Age	0.08	1.08	0.000
Chemotherapy (Yes)	0.94	2.57	0.001
Surgery (Mastectomy)	0.43	1.54	0.001
CTNNA1	0.49	1.63	0.005
CHEK2	-0.54	0.58	0.026
CCND3	0.42	1.53	0.028
NOTCH1	-0.39	0.68	0.033
FOXO3	0.31	1.36	0.049
(e) Luminal B Subtype			
Covariate	**Coefficient**	**HR**	**p value**
Age	0.04	1.04	0.000
Chemotherapy (Yes)	0.78	2.17	0.000
Surgery (Mastectomy)	0.36	1.44	0.001
CBFB	-0.46	0.63	0.008
CDKN1B	0.18	1.19	0.032
SETDB1	0.35	1.42	0.036
(f) Normal Subtype			
Covariate	**Coefficient**	**HR**	**p value**
Age	0.06	1.06	0.000
MAP2K4	-1.93	0.15	0.000
CTNNA1	1.47	4.36	0.000
RUNX1	-1.21	0.30	0.000
CDKN2A	-1.04	0.35	0.000
FOXO3	1.31	3.70	0.000
FOXO1	-1.57	0.21	0.001
MAP3K1	0.76	2.15	0.001
NRAS	1.22	3.38	0.001
EGFR	-0.35	0.70	0.001
ARID5B	1.12	3.07	0.004
ERBB3	-0.66	0.52	0.009
HRAS	0.86	2.37	0.011
JAK1	-0.70	0.50	0.019
TBX3	-0.41	0.66	0.034
CDH1	0.23	1.26	0.044

3.2. Survival Analysis

A comprehensive survival analysis was conducted where, the focus was placed on identifying optimal cutoff values for the expression levels of genes. Overall survival time and status were used to find the best cutoff points that could categorize patients into high and low expression groups for each gene. Kaplan-Meier survival curves were generated to visualize the survival probabilities of these groups, and the log-rank test was performed to assess the statistical significance ($p<0.05$) of the differences between them. This analysis revealed distinct high-risk and low-risk subgroups, highlighting the prognostic value of these genes across different breast cancer subtypes. The survival plots for the top two genes for each subtype based on Cox coefficient are presented in Figure 1. Survival Analysis (KM plot) of Basal-Like (a)-(b), Her2 (c)-(d), Claudin-Low (e)-(f), Luminal A (g)-(h), Luminal B (i)-(j) and Normal (k)-(l) subtypes.

3.3. PPI Network Analysis

The Protein-Protein interaction network is examined for 31 chosen protein-encoded genes that are found to be significantly correlated with breast cancer survival using the STRING database. An enrichment p-value of less than 1.0E-16 is found in the network. Every node in the network symbolizes a particular gene's relationship. In a similar vein, the edges represent relationships, when a group of related genes work together to support a common function. They don't have to engage in physical contact with one another, though. The PPI network and the node degrees of the high interaction genes are displayed in Figure 2. Hub proteins with greater node degrees, such as CDKN2A, AKT1, SMAD4, and NOTCH1, are observed, where AKT1 is an oncogene and SMAD4 and CDKN2A are tumour suppressor genes[14–16]. On the other hand, NOTCH1 is regarded as an

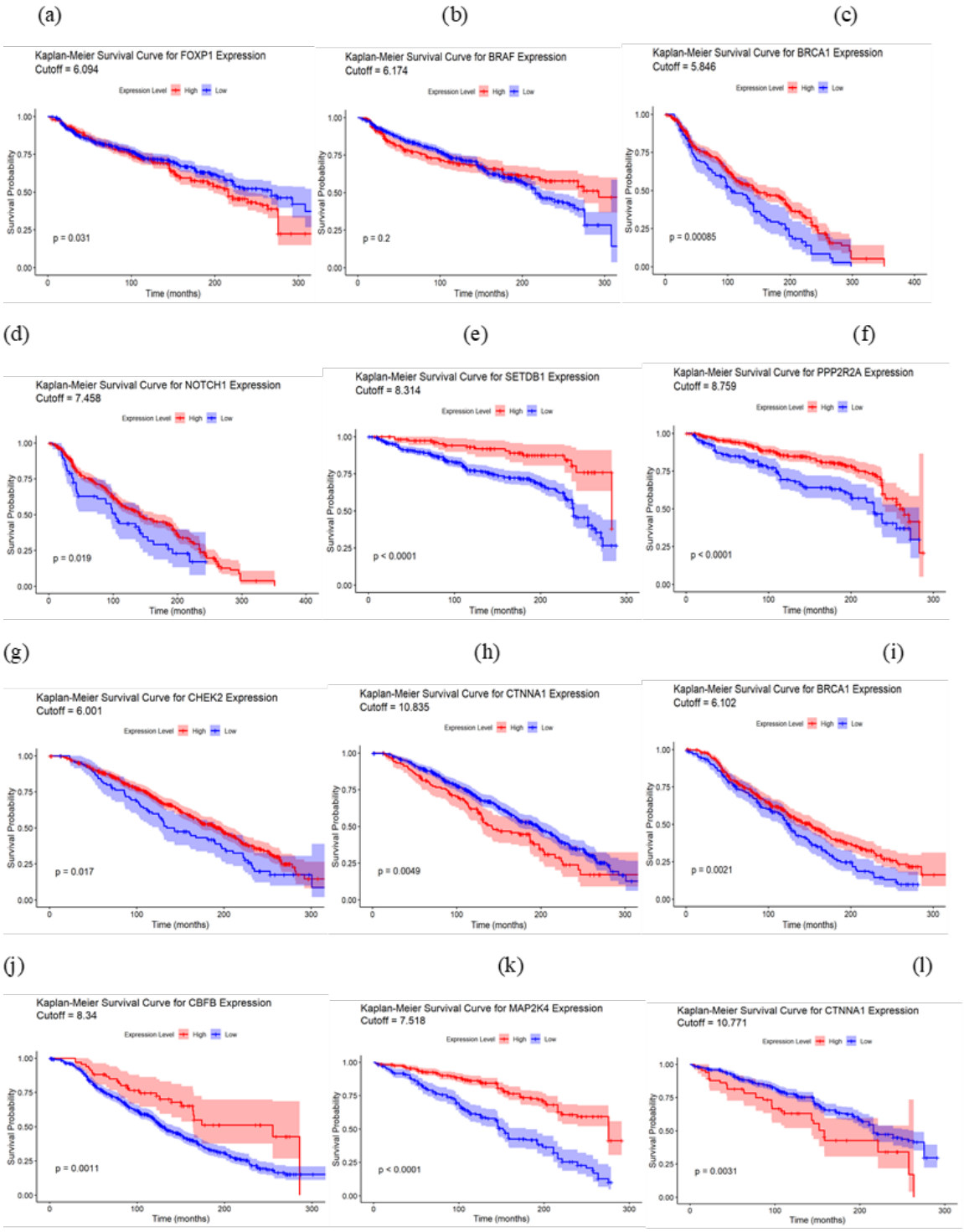
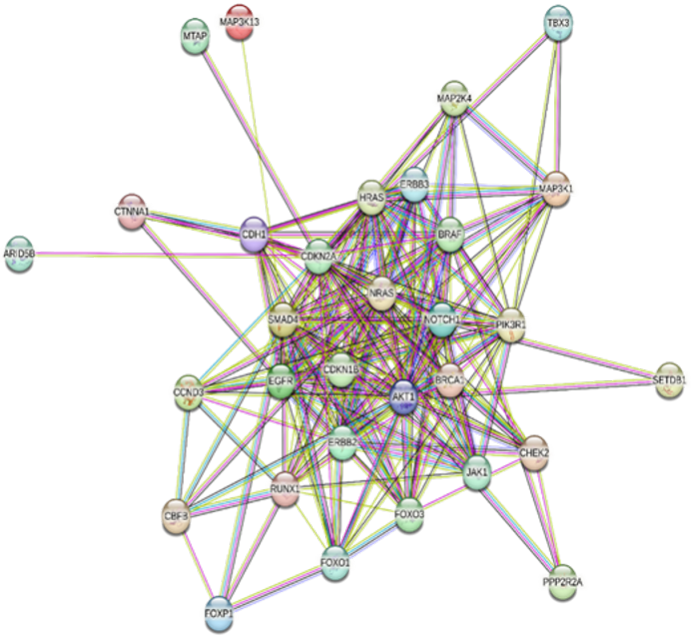


Fig 1. Survival Analysis (KM plot) of Basal-Like (a)-(b), Her2 (c)-(d), Claudin-Low (e)-(f), Luminal A (g)-(h), Luminal B (i)-(j) and Normal (k)-(l) subtypes

oncogene as well as a tumour suppressor gene[17]. As these genes are part of the breast cancer pathway, the analytical results indicate that the set of genes that have been found may be useful as biomarkers for the detection of breast cancer.



Name of Node	Node Degree
CDKN2A	24
AKT1	23
SMAD4	22
NOTCH1	21
CDKN1B	20
BRAF	19
BRCA1	19
NRAS	19
PIK3R1	18
EGFR	18
ERBB2	18
Average Node Degree	12.5
PPI Enrichment p-value	1.0E-16

Fig 2. Protein-Protein Interaction network of 31 genes

3.4. KEGG pathway analysis of selected genes

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was carried out with the SRPlots tool to reveal the pathway of the 31 important genes linked to the survival of breast cancer. Every pathway has a p-value, and a lower p-value (<0.05) indicates a greater likelihood of a pathway enriched with a certain set of genes. The top 10 KEGG pathways enriched for breast cancer, based on p-values, are displayed Figure 3. It revealed a notable enrichment of the PI3K-Akt signaling pathway. It is abnormally active in malignancies and plays a critical part in a number of cellular processes, which aid in the development and spread of tumors. Because of its potential for metastasis, it makes the cancer more aggressive and difficult to cure. Consequently, breast cancers with aberrant PI3K-Akt signaling tend to have a poorer prognosis[18]. Moreover, ErbB signaling pathway increases cell survival and proliferation, which can result in metastasis, and is a major factor in the aggressive behaviour of several breast cancer subtypes[19]. For the control of apoptosis, cell cycle arrest, and DNA repair in breast cancer, the FOXO signaling pathway, as identified by our KEGG pathway, is essential. When FOXO proteins' tumor-suppressive actions are inhibited, it can lead to the growth and metastasis of tumors[20]. These findings indicate that the identified genes play a crucial role in breast cancer.

3.5. GO Enrichment of the selected genes

The functional roles of the discovered genes were ascertained through the application of Gene Ontology (GO) analysis. By classifying the genes into distinct biological processes, cellular elements, and molecular roles, this research sheds light on how these genes influence the onset and spread of breast cancer. Figure 4 shows the top 10 enriched GO entries. GO terms associated with biological processes are positive regulation of protein serine/threonine kinase activity, negative regulation of cell adhesion, positive regulation of cyclin-dependent protein kinase activity etc. Among cellular components, breast cancer is also linked to significantly enriched GO keywords such as cell-cell junction and apical plasma membrane[21]). Moreover, protein tyrosine kinase activity, MAP kinase activity, and other items related to molecular functions are also significantly enriched GO terms[22].

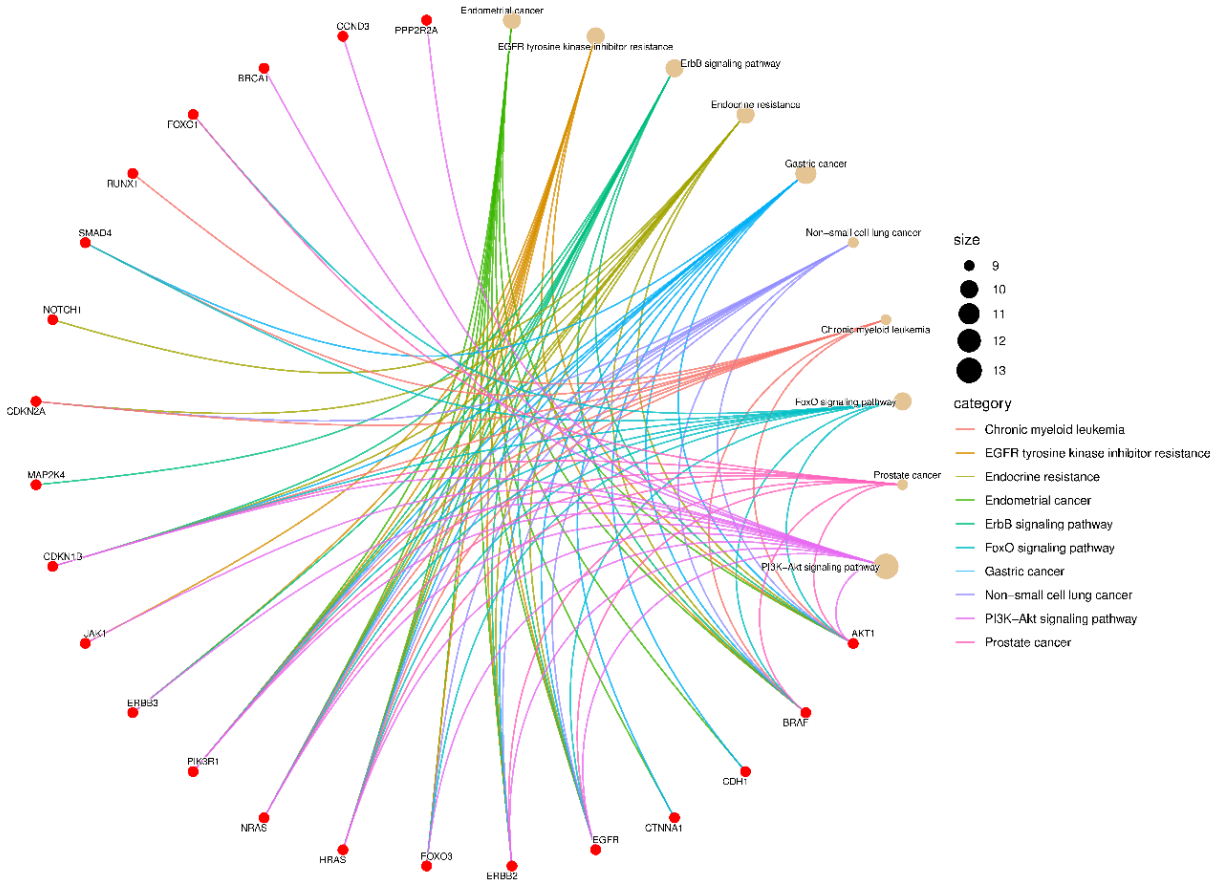


Fig 3. KEGG Pathways of the selected genes

4 Conclusion

Based on the studied dataset, LightGBM was identified as the most effective classifier for distinguishing breast cancer subtypes. Through Cox regression analysis, 31 significant genes impacting patient survival across different subtypes were pinpointed. Key findings are that higher expressions of genes like BRCA1 and NOTCH1 are associated with higher survivability. Whereas, overexpression of genes such as FOXP1 and CTNNA1 are linked to reduced survival. It was also discovered that ageing affected the survivability of all subtypes significantly.

Strong interactions were found between the identified genes according to protein-protein interaction (PPI) analysis, with the genes CDKN2A, AKT1, SMAD4, and NOTCH1 having the greatest node degrees, demonstrating their essential roles in the network. The KEGG Pathway analysis revealed a number of important pathways linked to these genes, including the pathways linked to endocrine resistance, different malignancies, and the PI3K-Akt signaling pathway. These pathways are known to be critical for the initiation and advancement of breast cancer. Important biological processes, cellular elements, and molecular functions that impact tumor growth, cell adhesion, hormone response and signaling have been revealed using Gene Ontology investigations.

The prognosis and course of treatment for breast cancer can significantly be affected by these findings. It is recommended that future studies verify these findings using various datasets to improve their reliability.

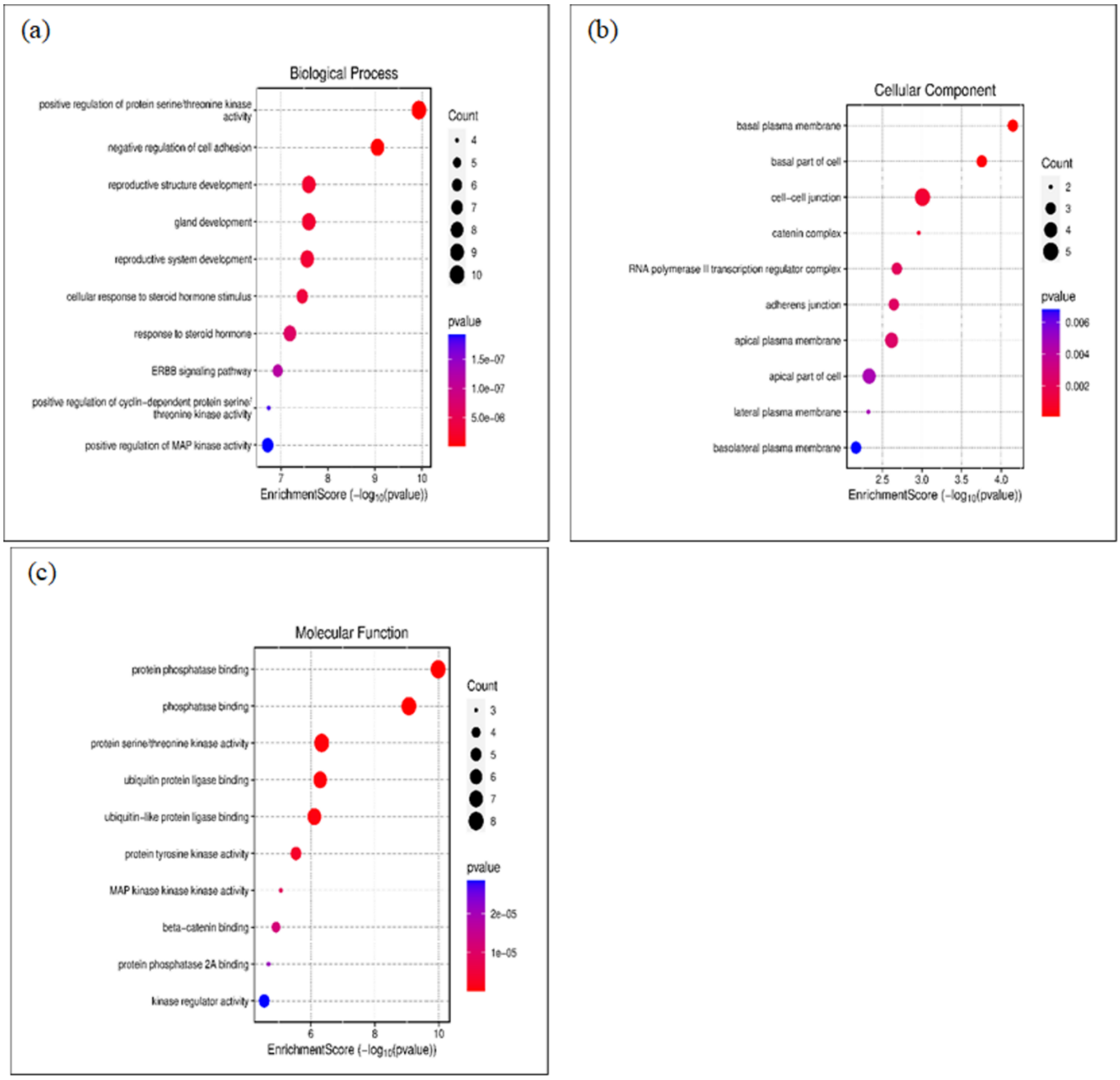


Fig 4. Gene Ontology analysis of the selected genes (a) Biological Process (b) Cellular Process (c) Molecular Process

Acknowledgment

The authors thank the anonymous reviewer and the editor for their invaluable feedback and suggestions, which significantly improved the quality of the manuscript. The authors also acknowledge cBioPortal, an open-source platform, for providing free access to high-quality datasets essential for this study.

References

1) Sarkar JP, Saha I, Sarkar A, Maulik U. Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. *Comput Biol Med [Internet]*. 2021;131. Available from: https://www.sciencedirect.com/science/article/pii/S001048252100038X.

2) Dehkordi FZS, Khatami SM, Ranjbar E. The Associations Between Urban Form and Major Non-communicable Diseases: a Systematic Review. *Journal of Urban Health* . 2022;99:941–58. Available from: https://doi.org/10.1007/s11524-022-00652-4.
3) Nenclares P, Harrington KJ. The biology of cancer. *Medicine [Internet]*. 2020;48(2):67–72. Available from: https://www.sciencedirect.com/science/article/pii/S1357303919302877.
4) D'arcy MS. Cell death: a review of the major forms of apoptosis, necrosis and autophagy. *Cell Biol Int*. 2019;43(6):582–92. Available from: https://doi.org/10.1002/cbin.11137.
5) Bashar MDA, Begam N. Breast cancer surpasses lung cancer as the most commonly diagnosed cancer worldwide. *Indian J Cancer [Internet]*. 2022;59(3). Available from: https://journals.lww.com/indianjcancer/fulltext/2022/59030/breast_cancer_surpasses_lung_cancer_as_the_most.26.aspx.
6) Deng JL, Xu Y, Hua, Wang G. Identification of Potential Crucial Genes and Key Pathways in Breast Cancer Using Bioinformatic Analysis. *Front Genet [Internet]*. 2019;10. Available from: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00695.
7) Aguilar D, Colacino JA. Single cell approaches to understand environmental impacts on aggressive breast cancers. *Curr Opin Toxicol [Internet]*. 2024;37:100459–100459. Available from: https://www.sciencedirect.com/science/article/pii/S2468202024000019.
8) Alhenawi E, Al-Sayyed R, Hudaib A, Mirjalili S. Feature selection methods on gene expression microarray data for cancer classification: A systematic review. *Comput Biol Med [Internet]*. 2022;140:105051–105051. Available from: https://www.sciencedirect.com/science/article/pii/S0010482521008453.
9) Zhao X, Lin J. Construction and Validation of a Prognostic Model Based on mRNAsi-Related Genes in Breast Cancer. *Comput Math Methods Med*. 2022;(1):6532591–6532591. Available from: https://doi.org/10.1155/2022/6532591.
10) Yang Y, Li Z, Zhong Q, Zhao L, Wang Y, Chi H. Identification and validation of a novel prognostic signature based on transcription factors in breast cancer by bioinformatics analysis. *Gland Surg* . 2022;11:892–912. Available from: http://dx.doi.org/10.21037/gs-22-267.
11) Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. . *Cancer Discov*. 2009;2:401–405. Available from: https://doi.org/10.1158/2159-8290.CD-12-0095.
12) Ghavidel A, Pazos P. Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: a systematic review. *Journal of Cancer Survivorship*. Available from: https://doi.org/10.1007/s11764-023-01465-3.
13) Fu X, Tan W, Song Q, Pei H, Li J. BRCA1 and breast cancer: Molecular mechanisms and therapeutic strategies. *Front Cell Dev Biol [Internet]*. 2022;10:813457–813457. Available from: http://dx.doi.org/10.3389/fcell.2022.813457.
14) George B, Gui B, Raguraman R, Paul AM, Nakshatri H, Pillai MR. AKT1 transcriptomic landscape in breast cancer cells. *Cells*. 2022;11(15):2290–2290. Available from: http://dx.doi.org/10.3390/cells11152290.
15) Zhao M, Mishra L, Deng CX. The role of TGF-β/SMAD4 signaling in cancer. *Int J Biol Sci*. 2018;14(2):111–134. Available from: https://www.ijbs.com/v14p0111.htm.
16) Cheng T, Wu Y, Liu Z, Yu Y, Sun S, Guo M. CDKN2A-mediated molecular subtypes characterize the hallmarks of tumor microenvironment and guide precision medicine in triple-negative breast cancer. . *Front Immunol*. 2022;13:970950–970950. Available from: http://dx.doi.org/10.3389/fimmu.2022.970950.
17) Edwards A, Brennan K. Notch signalling in breast development and cancer. *Front Cell Dev Biol [Internet]*. 2021;9:692173–692173. Available from: http://dx.doi.org/10.3389/fcell.2021.692173.
18) He Y, Sun MM, Zhang GG, Yang J, Chen KS, Xu WW. Targeting PI3K/Akt signal transduction for cancer therapy. *Signal Transduct Target Ther [Internet]*. 2021;6(1):425–425. Available from: https://doi.org/10.1038/s41392-021-00828-5.
19) Clusan L, Ferrière F, Flouriot G, Pakdel F. A basic review on estrogen receptor signaling pathways in breast cancer. *Int J Mol Sci [Internet]*. 2023;24(7). Available from: http://dx.doi.org/10.3390/ijms24076834.
20) Dilmac S, Kuscu N, Caner A, Yildirim S, Yoldas B, Farooqi AA, et al. SIRT1/FOXO Signaling Pathway in Breast Cancer Progression and Metastasis. *International Journal of Molecular Sciences*. 2022;23(18):10227–10227. Available from: https://dx.doi.org/10.3390/ijms231810227.
21) Mariscal LG, Miranda J, Gutiérrez HG, Cortina MC, Amaya E. Relationship between apical junction proteins, gene expression and cancer. *Biochimica et Biophysica Acta (BBA) - Biomembranes*. 2020;1862(9):183278–183278. Available from: https://dx.doi.org/10.1016/j.bbamem.2020.183278.
22) Galindo EG, Yilmaz ZH, Hausser A. Membrane trafficking in breast cancer progression: protein kinase D comes into play. *Frontiers in Cell and Developmental Biology*. 2023;11:1173387–1173387. Available from: http://dx.doi.org/10.3389/fcell.2023.1173387.