








Article

Identification of Novel Diagnostic and Prognostic Gene Signature Biomarkers for Breast Cancer Using Artificial Intelligence and Machine Learning Assisted Transcriptomics Analysis

Zeenat Mirza ^{1,2}, Md Shahid Ansari ³, Md Shahid Iqbal ⁴, Nesar Ahmad ⁴, Nofe Alganmi ^{5,6,7}, Haneen Banjar ^{5,6,7}, Mohammed H. Al-Qahtani ^{2,6} and Sajjad Karim ^{2,6,*}

- ¹ King Fahd Medical Research Center, King Abdulaziz University, Jeddah 21589, Saudi Arabia
 - ² Department of Medical Laboratory Science, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah 21589, Saudi Arabia
 - ³ Department of Clinical Data Analytics, Max Super Speciality Hospital, Saket, New Delhi 110017, India
 - ⁴ Department of Statistics and Computer Applications, Tilka Manjhi Bhagalpur University, Bhagalpur 812007, India
 - ⁵ Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
 - ⁶ Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia
 - ⁷ Centre of Artificial Intelligence in Precision Medicines, King Abdulaziz University, Jeddah 21589, Saudi Arabia
- * Correspondence: skarim1@kau.edu.sa; Tel.: +966-(55)-7581741



Citation: Mirza, Z.; Ansari, M.S.; Iqbal, M.S.; Ahmad, N.; Alganmi, N.; Banjar, H.; Al-Qahtani, M.H.; Karim, S. Identification of Novel Diagnostic and Prognostic Gene Signature Biomarkers for Breast Cancer Using Artificial Intelligence and Machine Learning Assisted Transcriptomics Analysis. *Cancers* **2023**, *15*, 3237. <https://doi.org/10.3390/cancers15123237>

Academic Editor: Patrizia Ferroni

Received: 15 May 2023
Revised: 10 June 2023
Accepted: 13 June 2023
Published: 18 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Simple Summary: Breast cancer is the most fatal female cancer, which the existing clinical and pathological information sometimes fails to diagnose accurately. Recent artificial intelligence-based studies have shown the capability of identifying molecular biomarkers using high-throughput genomics data. Our aim was to apply machine learning methods to a large cohort of transcriptomics data for gene reduction and the construction of a diagnostic model for cancer classification. Advanced statistical methods and cross-validation with another set of machine learning methods increased the accuracy of the diagnostic model and predicted a novel diagnostic nine-gene signature. Further, survival analysis revealed a novel prognostic model of eight-gene signatures. Experimental validation confirmed the expression of the identified gene signatures in breast cancer patients and increased the reliability of the study. The identified gene signature biomarkers have the potential to improve healthcare management with precise diagnosis and prognosis at a reduced cost.

Abstract: Background: Breast cancer (BC) is one of the most common female cancers. Clinical and histopathological information is collectively used for diagnosis, but is often not precise. We applied machine learning (ML) methods to identify the valuable gene signature model based on differentially expressed genes (DEGs) for BC diagnosis and prognosis. Methods: A cohort of 701 samples from 11 GEO BC microarray datasets was used for the identification of significant DEGs. Seven ML methods, including RFECV-LR, RFECV-SVM, LR-L1, SVC-L1, RF, and Extra-Trees were applied for gene reduction and the construction of a diagnostic model for cancer classification. Kaplan–Meier survival analysis was performed for prognostic signature construction. The potential biomarkers were confirmed via qRT-PCR and validated by another set of ML methods including GBDT, XGBoost, AdaBoost, KNN, and MLP. Results: We identified 355 DEGs and predicted BC-associated pathways, including kinetochore metaphase signaling, PTEN, senescence, and phagosome-formation pathways. A hub of 28 DEGs and a novel diagnostic nine-gene signature (*COL10A*, *S100P*, *ADAMTS5*, *WISP1*, *COMP*, *CXCL10*, *LYVE1*, *COL11A1*, and *INHBA*) were identified using stringent filter conditions. Similarly, a novel prognostic model consisting of eight-gene signatures (*CCNE2*, *NUSAP1*, *TPX2*, *S100P*, *ITM2A*, *LIFR*, *TNXA*, and *ZBTB16*) was also identified using disease-free survival and overall survival analysis. Gene signatures were validated by another set of ML methods. Finally, qRT-PCR results confirmed the expression of the identified gene signatures in BC. Conclusion: The ML

approach helped construct novel diagnostic and prognostic models based on the expression profiling of BC. The identified nine-gene signature and eight-gene signatures showed excellent potential in BC diagnosis and prognosis, respectively.

Keywords: breast cancer; gene expression profiling; artificial intelligence; machine learning methods; diagnostic and prognostic model

1. Introduction

Breast cancer (BC) is the most common cause of cancer death in women, with 1 in 8 cancer cases, and its incidence has increased significantly despite the preventive and curative approaches utilized in recent years [1,2]. In 2020, the International Agency for Research on Cancer (IARC) and World Health Organization (WHO) reported 2.26 million new BC cases and 684,996 global cases of BC mortality in females, surpassing lung cancer with 2.20 million new cases. Further, the diagnosis of new cases and BC death by 2040 is predicted to increase to over 3 million and 1 million, respectively [3,4]. BC is a heterogeneous disease and the symptoms may include a bump, skin dimpling, nipple discharge, scaly hair patch and flaky skin around the nipple, and thickness/swelling in some parts of the breast [5]. BC survival rates were found to be variable, at ~80%, ~60%, and ~40% in high-, mid-, and low-income countries, respectively.

An accurate diagnosis is key for the optimal treatment of cancer patients. At present, cancer classification and diagnosis heavily depend on the subjective evaluation of physical examination, clinical/pathological test, radiological scan, and histopathological information, but they are subject to human errors [6]. Surprisingly, medical error is the third leading cause of death, even in the most advanced countries such as the USA [7]. Additionally, in some instances, (i) incomplete or misleading clinical information, (ii) complicated radiological images, and (iii) variable, atypical, or lack of morphologic features in histological information may result in diagnostic confusion, and thus affect patient care [2].

Molecular diagnostics offer precise, fair, and efficient breast cancer classification, but are not widely applied in clinical settings. Microarray-platform-based assays, including the Affymetrix GeneChip Human Genome U133 Plus 2.0 array (Affymetrix, Santa Clara, CA, USA), have the ability to measure thousands of gene expressions simultaneously for each data point (sample) [8,9]. Expression profiling to check for variability in gene expression is an important factor influencing the precision and accuracy of clinical decisions in the diagnosis of BC [1,8,10–13]. Despite the large-scale, high-dimensional, and highly redundant type of microarray data, with numerous tools to identify genes that are differentially expressed across cancer/disease phenotypes, the interpretation of results and follow-up analysis are quite challenging. DNA-microarray-based gene expression profiling is promising for BC diagnosis and prognosis [12,14], but limitations such as small sample size, biased case vs. control distribution, multiple BC subtypes, variable populations, and different platforms, complicate the analysis, and the identification of gene signatures remains an issue [15–19].

In 1999, the first gene expression signature was identified to classify leukemia into acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Since then, a series of gene expression signatures have been reported for various cancers to classify tumors, tumor types, tumor stages, and predict the disease prognosis [20,21].

BC is a very heterogeneous disease and is categorized into five molecular subtypes: HER2+, basal (ER−/HER2−/PR−), luminal A (ER+/HER2, with a low-proliferative phenotype), luminal B (ER+/HER2, with a high-proliferative phenotype), and normal BC [17]. Each subtype exhibits distinct transcriptomics patterns, and finding unified BC biomarkers or the gene signature applicable to all molecular subtypes remains a challenge [13,22]. Hence, multiple datasets need to be integrated to find universal diagnostic/prognostic biomarkers broadly applicable to all BC subtypes.

A stringent filtration condition drastically reduces the number of DEGs, but filters out many biologically relevant genes as well, whereas a lenient cutoff allows for many genes to pass through, and a follow-up issue arises in selecting the most interesting genes. Although conducting gene ontology and pathway enrichment analysis is useful in predicting biological processes, cellular components, molecular function, networks, and canonical pathways for the detected DEGs, the selection of the most relevant genes, a diagnostic and/or prognostic biomarker, in cancer remains a challenge. Machine learning methods and different evaluation techniques such as the Kaplan–Meier (KM) estimator might be useful in identifying the biologically relevant genes from a long DEG list, without any obvious selection way [8,11].

Another problem in high-throughput gene expression profiling is reducing the extremely high dimensionality of irrelevant or redundant gene features responsible for cancer classification accuracy. Feature selection methods have been used to select key genes from thousands of expressed genes, but the large numbers of microarray genes used in most existing methods for cancer classification often hamper the model outcomes. For an efficient diagnostic model for BC, machine-learning-based feature selection methods were applied to a smaller number of differentially expressed genes passing the standard statistical cutoff of $p < 0.5$ and \log_2 folds change > 2 in BC.

To address these challenges, we integrated eleven GEO oligonucleotide microarray datasets to create a gene expression database of 701 samples (356 breast tumors and 345 normal breast tissues) and applied different R packages and machine learning methods on gene expression data for the molecular classification, accurate diagnosis, and prognostic evaluation of the identified gene signatures in BC. A larger sample size gave greater analysis power because it constricted the distribution of the test statistic. Further, an almost equal group (3356 vs. 3345) reduced the biases in machine-learning-based data analysis, and increases the accuracy of the model and predicted biomarkers. We also demonstrated that the combined use of molecular pathway analysis, expression analysis, feature selection methods, and survival analysis was helpful in selecting gene signatures with high confidence.

2. Materials and Methods

2.1. Data Sets and Patients

The raw gene expression data, a set of binary files in a CEL format, of BC from eleven datasets, including GSE61304, GSE42568, GSE7904, GSE3744, GSE29431, GSE26910, GSE31138, GSE71053, GSE10780, GSE30010, and GSE111662, were retrieved from the Gene Expression Omnibus (GEO) database using “GEOquery” library of the R program (<https://www.ncbi.nlm.nih.gov/geo/>, accessed on 2 January 2023). We selected only the human breast tumor samples to eliminate differential genetic interference in different BC cell lines. Clinicopathological information from the original studies was used for analysis. The ratio of breast tumors to normal breast was biased in the majority of the deposited GEO datasets, including GSE61304, GSE42568, GSE7904, GSE3744, and GSE26910. Thus, we included additional normal breast cases (GSE30010 and GSE111662) to balance the data ($n = 701$, $356 = \text{BT}$ vs. $345 = \text{NB}$) for a better outcome, while identifying DEGs or applying ML methods to develop a diagnostic model (Table 1). This study was approved by the university’s CEGMR bioethical committee (16-CEGMR-bioeth-2022), dated 13 October 2022, and we recruited patients for the validation of potential biomarkers after obtaining their consent.

2.2. Preprocessing and Differential Expression Analysis

The median expression values of less than 5.55 intensity on the \log_2 scale of each probe, indicating the failure of true hybridization, were filtered out. We also excluded the probes expressed in less than two samples. We merged all the raw CEL files ($n = 701$) and applied the RMA method for the normalization of expression values, and generated box plots using the “oligo” package from R software. Principal component analysis (PCA) was

performed using “prcomp function”, and hierarchical clustering was carried out using the “pheatmap” R package to correlate the samples with the probes.

Table 1. GEO datasets from the GPL570 platform used for gene expression profiling.

| Dataset | Title/Description | Normalization Methods | No. of Samples | Percentage of Cancer |
|-----------|--|---|--|----------------------|
| GSE61304 | Novel biomarker discovery for stratification and prognosis of breast cancer patients | MAS5 signal intensity | 62 (58 breast tumor + 4 normal breast) | 94% |
| GSE42568 | Breast cancer gene expression analysis | Log2 GCRMA signal intensity | 121 (104 breast tumor + 17 normal breast) | 86% |
| GSE7904 | Expression data from human breast tissue | RMA expression value | 50 (43 breast tumor + 7 normal breast) | 86% |
| GSE3744 | Human breast tumor expression | GCRMA calculated signal intensity, log2 transformed | 47 (40 breast tumor + 7 normal breast) | 85% |
| GSE29431 | Identifying breast cancer biomarkers | RMA expression values | 66 (54 breast tumor + 12 normal breast) | 82% |
| GSE26910 | Stromal molecular signatures of breast and prostate cancer | Log2 RMA signal | 12 (6 breast tumor + 6 normal breast) | 50% |
| GSE31138 | Identifying novel anti-angiogenic targets in human breast cancer | Log2 RMA signal | 6 (3 breast tumor + 3 normal breast) | 50% |
| GSE71053 | Differential effect of surgical manipulation on gene expression in normal breast tissue and breast tumour tissue | Log2-normalized signal | 18 (6 breast tumor + 12 normal breast) | 33% |
| GSE10780 | Proliferative genes dominate malignancy risk gene signature in histologically normal breast tissue | RMA expression value | 185 (42 breast tumor + 143 normal breast) | 23% |
| GSE30010 | Expression data from breast samples of postmenopausal women | RMA expression value | 107 (0 breast tumor + 107 normal breast) | 0% |
| GSE111662 | Whole breast tissue gene expression in comparison to expression in epithelial and stromal tissues | RMA expression values | 27 (0 breast tumor + 27 normal breast) | 0% |
| Total | | | 701 (356 breast tumors + 345 normal breasts) | 51% |

We used linear models for the microarray “Limma” package of R to identify differentially expressed genes (DEGs), using an empirical Bayesian method to assess the differences in gene expression. Wang et al. (2021) conducted a study demonstrating the superior performance of the moderated *t*-test when the sample size was ≥ 40 . [23,24]. Our study, with a sample size of 701, comprising 356 breast tumors and 345 normal breast samples, exceeds the required 95% power of the test, and the nearly equal representation of the test (tumor) and control (normal) groups mitigates biases in machine-learning-based data analysis, and enhances the accuracy of the model and predicted biomarkers. The “decideTests” function was used to differentiate between the altered (up or down) and normal expression. The “topTable” function from R was applied with a cut-off-adjusted *p*-value (Benjamini–Hochberg-corrected false discovery rate) < 0.05 and log2 fold change $> \pm 2$ to detect the most significant DEGs in BC compared with normal samples. Unannotated

probes, not representing genes, were removed, and duplicate probes, representing single genes, were averaged for expression values to get a unique set of DEGs.

2.3. Functional Pathway and Gene Set Enrichment Analysis

We used a comprehensive set of functional annotation tools, such as the QIAGEN Ingenuity Pathway Analysis (IPA knowledgebase v84978992, QIAGEN, USA) and WEB-based Gene SeT AnaLysis Toolkit (WebGestalt 2019, <https://www.webgestalt.org/>, accessed on 2 January 2023), to investigate and understand the biological meaning of long-list significant DEGs [1,25,26]. We explored gene ontologies, enriched and canonical pathways, upstream regulators, disease and functions, and the networks associated with BC. Over-representation (or enrichment) analysis (ORA), a statistical method, was used to determine the presence of known genes in pre-defined sets, as well as in dataset/DEGs.

2.4. Machine Learning and Feature Selection Methods

We applied machine learning methods to BC transcriptomics data and considered performance measurements such as classification accuracy, specificity, sensitivity, and AUC, to identify the most informative features. To determine the effectiveness of a classification model, a set of performance metrics was used for assessment, such as measuring the model's ability to accurately classify instances into the correct categories. We used the confusion matrix to compute the accuracy, precision, recall, and F1 score, as shown in Equations (1)–(4).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{\sum TP}{\sum TP + \sum FP} \quad (2)$$

$$Recall = \frac{\sum TP}{\sum TP + \sum FN} \quad (3)$$

$$F1 - score = \frac{2 * precision * recall}{precision + recall} \quad (4)$$

In addition, the model's performance was evaluated by plotting the receiver operating characteristic curve (ROC) and the area under the ROC (AUC), which is a metric used to measure the model's effectiveness. Models with larger AUCs are considered to have higher performance.

The Scikit-learn (sklearn) in python platform was used to build the ROC curves of the DEGs and measure the AUC to compare the diagnostic value of the DEGs, and to predict the accuracy of the detected DEGs. The ROC curve, reflecting the relationship between sensitivity and specificity, and AUC were used to determine the diagnostic value of a factor in a specific disease, with AUC values between 0.5 and 1 representing low and high authenticity, respectively.

We used seven machine learning algorithms, including (i) recursive feature elimination with cross validation (RFECV) with logistic regression, (ii) RFECV with support vector machine (SVM), (iii) Lasso regularization (L1) with logistic regression, (iv) Lasso regularization (L1) with support vector classification (SVC-L1), (v) random forest (RF) classifier, (vi) extremely randomized trees (extra trees) classifier, and (vii) genetic algorithms (GA), to find the most significantly expressed genes in all samples (n = 701, 356BT + 345NB) and construct the diagnostic model from candidate DEGs. To validate the constructed diagnostic and prognostic models, we used five additional ML methods, including (i) adaptive boosting (AdaBoost), (ii) gradient-boosted decision trees (GBDT), (iii) K-nearest neighbors (KNN), (iv) multilayer perceptron (MLP), and (v) the extreme gradient boosting (XGBoost).

2.4.1. RFECV with Logistic Regression or with SVM

We used RFECV with logistic regression and SVM in Python using the RFECV class from the scikit-learn library. The RFECV with logistic regression or SVM is a method of feature selection in machine learning. It is a combination of two techniques: recursive feature elimination (RFE) and cross validation (CV). RFE is a backward selection algorithm that starts with all the features and removes the weakest feature until a specified number of features is left. CV, on the other hand, is a technique used to evaluate the performance of a model by dividing the data into several folds and training the model on different folds, while testing it on one fold at a time. In RFECV with logistic regression or SVM, the RFE algorithm is combined with CV to eliminate features, while also evaluating the performance of the logistic regression model. This helps determine the optimal number of features that provide the best performance, while avoiding overfitting. By combining these two techniques, RFECV with logistic regression or SVM ensures that the final feature set is not only informative, but also generalizable to new data.

2.4.2. LASSO Regularization (L1) Using Logistic Regression or Support Vector Classification

We used L1 with logistic regression and SVM from the scikit-learn library to identify the most significant genes [27]. LASSO regularization is a technique used to reduce the number of features in a model and prevent overfitting. The technique shrinks the magnitude of the coefficients using a penalty term proportional to the absolute value of the coefficients, resulting in some coefficients becoming zero. It helps select the most relevant features, while reducing the impact of irrelevant or noisy features on the model's performance. L1 regularization is particularly useful when dealing with a large number of features or highly correlated features. The loss function of Lasso regression is defined as shown in (5):

$$\sum_{i=1}^n (Y_i - \sum_{j=1}^p X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (5)$$

where lambda is the regularization parameter that controls the strength of the penalty term.

In logistic regression with LASSO regularization, the L1 penalty term helps reduce the impact of irrelevant or noisy features on the model's performance by shrinking their coefficients toward zero. This can improve the model's interpretability and reduce the risk of overfitting, especially when dealing with high-dimensional data.

In SVM, LASSO regularization is implemented by adding a penalty term to the objective function that minimizes the classification error. The penalty term is dependent on the magnitude of the coefficients of the features, so larger coefficients receive a larger penalty. This ensures that features with a large impact on the classification result receive a smaller penalty and are more likely to be included in the final model.

2.4.3. Random Forest

The random forest classifier is a machine learning algorithm used for both classification and regression problems [28]. It is an ensemble of decision trees, where each tree is trained on a random subset of data. The final prediction is made by taking the average of all the trees' predictions. In feature selection, a random forest classifier is used to select the most important features in the dataset. The algorithm calculates the importance of each feature by measuring the average decrease in impurity for that feature. The higher the average decrease, the more important the feature is considered.

2.4.4. Extra Trees Classifier

The extra trees classifier is an ensemble machine learning algorithm that can be used for feature selection in Python [29]. It is a type of random forest classifier where multiple decision trees are grown and combined to make a prediction. The algorithm works by randomly selecting a subset of features at each split in the tree and determining the most

important features based on their impact on the final prediction. By aggregating the feature importance scores across all trees, the extra trees classifier can provide a ranking of the most important features for a given dataset. This can be useful in identifying the most relevant features for building a predictive model and reducing the dimensionality of the data.

2.4.5. Genetic Algorithm

This algorithm uses principles of evolution and natural selection to find the optimal combination of features that result in the best model performance. The algorithm starts with a random set of features and uses a fitness function to evaluate the performance of each combination. The best performing combinations are then recombined and mutated to create a new generation of features, and the process continues until a satisfactory set of features is found [30].

2.4.6. XGBoost

This algorithm is an ensemble learning method that works by combining multiple weak models into a strong one [31]. It uses gradient boosting, which is a method of iteratively training decision trees on residuals to improve the model performance. It provides faster computation and parallelization of training, which is useful when working with large datasets. It also has built-in regularization techniques to reduce overfitting, which is a common problem in machine learning.

2.4.7. GBDT

This is a machine learning algorithm that works by building an ensemble of decision trees in a way that each subsequent tree focuses on the errors made by the previous trees [32]. This iterative process results in a model that can learn complex non-linear relationships in data. It has the ability to handle large datasets, handle missing data, and provide accurate predictions with high interpretability.

2.4.8. MLP

This is a type of feedforward artificial neural network that consists of multiple layers of nodes that process information from the input layer to the output layer through a series of nonlinear transformations [33]. The nodes in each layer are connected to the nodes in the previous and next layers, and each node applies an activation function to the weighted sum of its inputs. The goal of training the MLP is to minimize this cost function by adjusting the weights and biases in the network using an optimization algorithm such as gradient descent. This allows the model to learn the best set of weights that can accurately predict the binary classification labels for unseen data. The cost function in the binary classification of MLP uses the binary cross-entropy loss of function and is defined as shown in (6):

$$\text{cross-entropy} = -\frac{1}{N} \sum_{i=1}^N (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \quad (6)$$

where N is the number of samples,

y_i is the actual outcome,

p_i is the probability of the tumor class, and

$1 - p_i$ is the probability of the normal class.

2.4.9. AdaBoost

This is an ensemble learning algorithm that combines several weak learners to create a strong learner [34]. It works by repeatedly fitting a weak learner to the data, and adjusting the weights of the training samples to focus on the misclassified ones. The algorithm then combines these weak learners to form a strong learner that is capable of accurately predicting the target variable.

2.4.10. KNN

The K-nearest neighbors (KNN) is a popular machine learning algorithm belonging to the family of instance-based or lazy learning algorithms, which means that it does not attempt to learn a function from the training data [35]. Instead, KNN stores all the training data, and classifies new data based on the similarity of its features to those in the training set. The number of nearest neighbors (K) is a hyperparameter that can be tuned to improve performance.

2.5. Survival Analysis Using the Kaplan–Meier Estimator

The KM estimator, a statistical technique tool (available at <https://kmplot.com/analysis/>, accessed on 10 February 2023) was used for calculating survival probability functions to investigate the overall and relapse-free survival of prognostic genes for breast cancer patients. It is assumed that the occurrence of the event is fixed in time, and both the censored observations and data points have an equal chance of survival [8,36].

The mathematical expression of KM is expressed as shown in (7):

$$S(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \quad (7)$$

$S(t)$ stands for survival function.

In this context, n_i refers to the count of individuals at risk at a specific time t_i , and d_i is the count of events that happen at the same time, t_i . The survival curve remains unchanging between the two events or times, i.e., between t_i and $t_i + 1$ [36].

This analysis was conducted for mRNA (gene chip) microarray data for the relapse-free and overall survival. The KM analysis was performed with a confidence interval and log rank p -value cut-off of $>95\%$ and ≤ 0.05 , respectively. We proceeded to further check the mRNA (RNA seq) datasets for overall survival for genes which were significant in both microarray data for relapse-free survival and overall survival. Finally, we established eight gene hubs (four upregulated and four downregulated) for prognostic importance.

The web-based KMplot tool incorporates three databases in the background: TCGA, EGA, and GEO [37]. The Kaplan–Meier method is a strong non-parametric statistical approach used for predicting the likelihood of survival. The KM analysis was performed with a confidence interval and log rank p -value cut-off of $>95\%$ and ≤ 0.05 , respectively.

2.6. RNA Isolation and qRT-PCR

Trizol was used to lyse the cells, and chloroform and isopropanol were used to extract RNA. After determining the RNA concentration, the cDNA (complimentary deoxyribonucleic acid) was reverse-transcribed. The primer sets were designed for the identified gene signature using Primer-3 software (V.0.4.0). ABI 7500 instruments were used for real-time quantitative PCR. Endogenous *GAPDH* gene expression was measured as the internal control to determine the relative expression of the detected genes. The reaction was run in a final volume of 10 μ L, comprising 5 μ L SYBR-Green qPCR master mix (KAPA Biosystems, Wilmington, MA, USA), 10 pmol of each primer, and 20 ng genomic DNA. PCR was performed in triplicate using the SYBR-Green qPCR master mix (KAPA Biosystems, USA) in a 96-well plate. Raw data were generated through the use of StepOne Plus™ Real-Time PCR Systems and Data Assist software. qPCR data were analyzed by $\Delta\Delta$ CT or the Livak method, and the GraphPad PRISM software was used for presentation.

2.7. Statistical Analysis

All statistical analyses were conducted using R software (version v.4.2.2) (R core team 2021). R was also used for the picture generation. The chi-squared test was used to compare categorical variables of patient characteristics. The Wilcoxon rank sum test was used to compare the expression signature. The p -values were adjusted for multiple comparisons using the Benjamini–Hochberg method, and the default value <0.05 was considered statistically

significant, otherwise specified. Cox regression analysis (univariate and/or multivariate) was used to assess the contribution of all parameters, such as evaluating the independent predictive OS performance of different clinical factors and the detected biomarkers. The KM curve and time-dependent ROC curve were drawn by the R package “survminer” and “survivalROC”, respectively.

3. Results

3.1. Differentially Expressed Genes in BC

Breast tumor (356) and normal breast tissue (345) samples from the Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays platform (HG-U133_Plus_2) with 54,675 features/probes from 11 different GEO data series were used as discovery cohorts. Out of 54,675 hybridized probes, only 46,597 probes passed the cut-off: median expression >5.5 and present in at least two samples. The expression data were GC-RMA-normalized. The raw intensities and RMA-normalized expression values were shown in Boxplot (Figure 1).

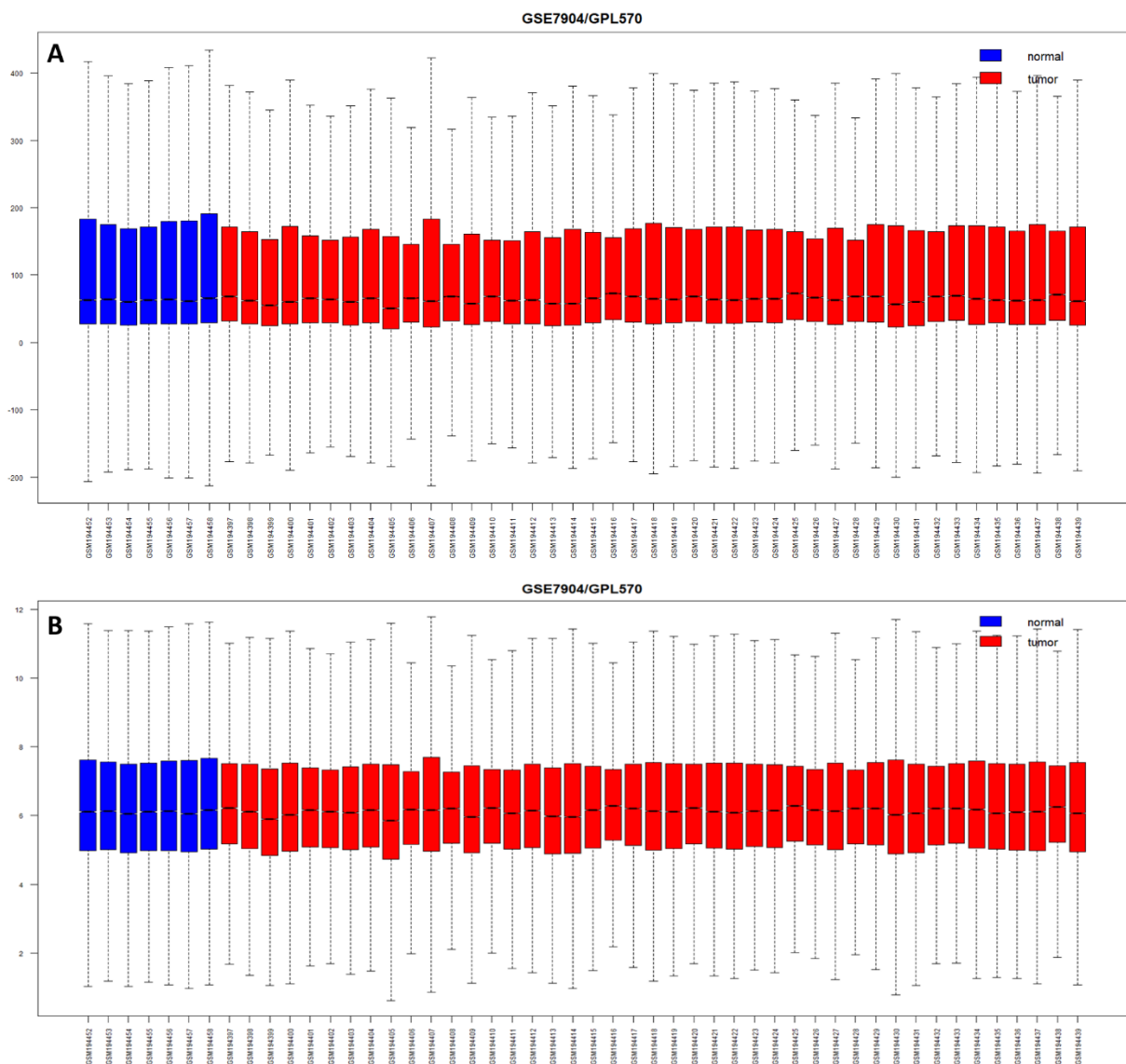


Figure 1. Boxplot showing the expression distribution for dataset GSE7904. (A) Raw (un-normalized) expression distribution with log scale in the range of -200 to 400 . (B) Normalized intensities showing almost similar distributions of expression intensities, with the log₂ scale in the range of 0 to 12 .

For a tumor-normal matrix, the “decideTests” function differentiated 46,597 probe signal intensities into altered (over-expressed (15,000), under-expressed (21,952)) and unaltered (9645) expression. The “topTable” functions revealed the most significant DEGs in breast cancer (n = 487) for the screening criteria of the adjusted P-value (Benjamini–Hochberg-corrected false discovery rate) < 0.05 and fold change ($\log_2FC > \pm 2$). Additionally, unannotated probes, not representing valid genes (n = 20), were removed first, then duplicate genes, multiple probes representing single genes (n = 193), were averaged for expression values (n = 83) to get a unique set of DEGs (n = 355, upregulated = 77 and downregulated = 278) (Figure 2 and Table 2).

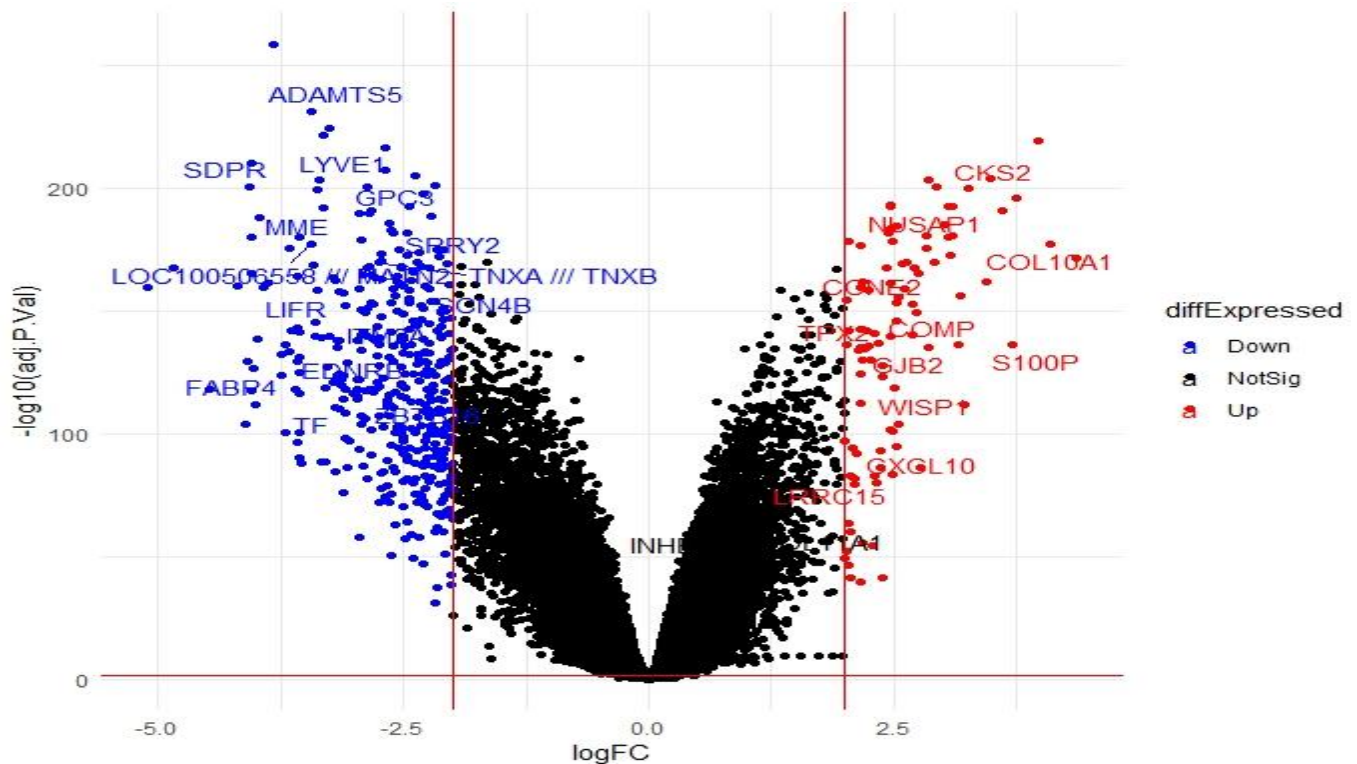


Figure 2. Volcano plot showing differentially expressed genes: (i) the majority were non-significant (black), (ii) upregulated DEGs (red), and (iii) downregulated DEGs (blue).

Table 2. Top ten up- and downregulated differentially expressed genes from a total of 355 DEGs in breast cancer.

| Gene Symbol | Gene Name | Log ₂ FC | adj. <i>p</i> -Value | Decide Test |
|----------------|--|---------------------|-------------------------|-------------|
| <i>COL11A1</i> | Collagen Type XI Alpha 1 Chain | 4.36 | 1.69×10^{-172} | Upregulated |
| <i>TOP2A</i> | DNA Topoisomerase II Alpha | 3.96 | 4.16×10^{-220} | Upregulated |
| <i>S100P</i> | S100 Calcium-Binding Protein P | 3.70 | 3.57×10^{-137} | Upregulated |
| <i>COL10A1</i> | Collagen Type X Alpha 1 Chain | 3.59 | 7.47×10^{-192} | Upregulated |
| <i>RRM2</i> | Ribonucleotide Reductase Regulatory Subunit M2 | 3.47 | 4.44×10^{-205} | Upregulated |
| <i>CKS2</i> | CDC28 Protein Kinase Regulatory Subunit 2 | 3.26 | 8.66×10^{-201} | Upregulated |
| <i>MMP1</i> | Matrix Metalloproteinase 1 | 3.21 | 7.95×10^{-113} | Upregulated |
| <i>COMP</i> | Cartilage Oligomeric Matrix Protein | 3.15 | 6.27×10^{-137} | Upregulated |
| <i>NUSAP1</i> | Nucleolar And Spindle-Associated Protein 1 | 3.08 | 9.67×10^{-194} | Upregulated |

Table 2. Cont.

| Gene Symbol | Gene Name | Log2FC | adj.-p-Value | Decide Test |
|-------------|--|--------|-------------------------|---------------|
| ANLN | Anillin, Actin-Binding Protein | 3.07 | 2.42×10^{-173} | Upregulated |
| ADH1B | Alcohol Dehydrogenase 1B (Class I), Beta Polypeptide | -4.84 | 2.10×10^{-168} | Downregulated |
| ADIPOQ | Adiponectin, C1Q And Collagen Domain Containing | -4.47 | 6.08×10^{-119} | Downregulated |
| PLIN1 | Perilipin 1 | -4.20 | 8.42×10^{-161} | Downregulated |
| LEP | Leptin | -4.11 | 7.20×10^{-105} | Downregulated |
| LPL | Lipoprotein Lipase | -4.09 | 2.35×10^{-130} | Downregulated |
| SDPR | Serum Deprivation Response | -4.06 | 1.34×10^{-201} | Downregulated |
| RBP4 | Retinol Binding Protein 4, Plasma | -4.06 | 4.37×10^{-118} | Downregulated |
| C2orf40 | Chromosome 2 Open Reading Frame 40 | -4.05 | 3.15×10^{-211} | Downregulated |
| ABCA8 | Atp-Binding Cassette Subfamily A Member 8 | -4.05 | 4.21×10^{-166} | Downregulated |
| NTRK2 | Neurotrophic Tyrosine Kinase, Receptor, Type 2 | -4.04 | 4.78×10^{-181} | Downregulated |

3.2. Function Pathway Analysis and Network Enrichment Analysis

Functional analysis based on the Z-score and $-\log(p\text{-value})$ indicated activation of the kinetochore metaphase signaling pathway (2.71, 7.35), PTEN pathway (2.64, 2.31), HOTAIR regulatory pathway (2.12, 2.71), and WNT/ β -catenin signaling pathway (2.45, 1.47), and suppression of the senescence pathway (-3.16, 3.02), phagosome formation (-3.15, 2.09), FAK signaling (-3.128, 1.74), oxytocin signaling pathway (-3.05, 3.8), and breast cancer regulation by Stathmin1 (-3.0, 2.06) (Figure 3). Most significantly enriched molecular processes were the extracellular matrix, cell division, mitotic cell cycle process, cell migration, and the regulation of cell proliferation (Table 3).

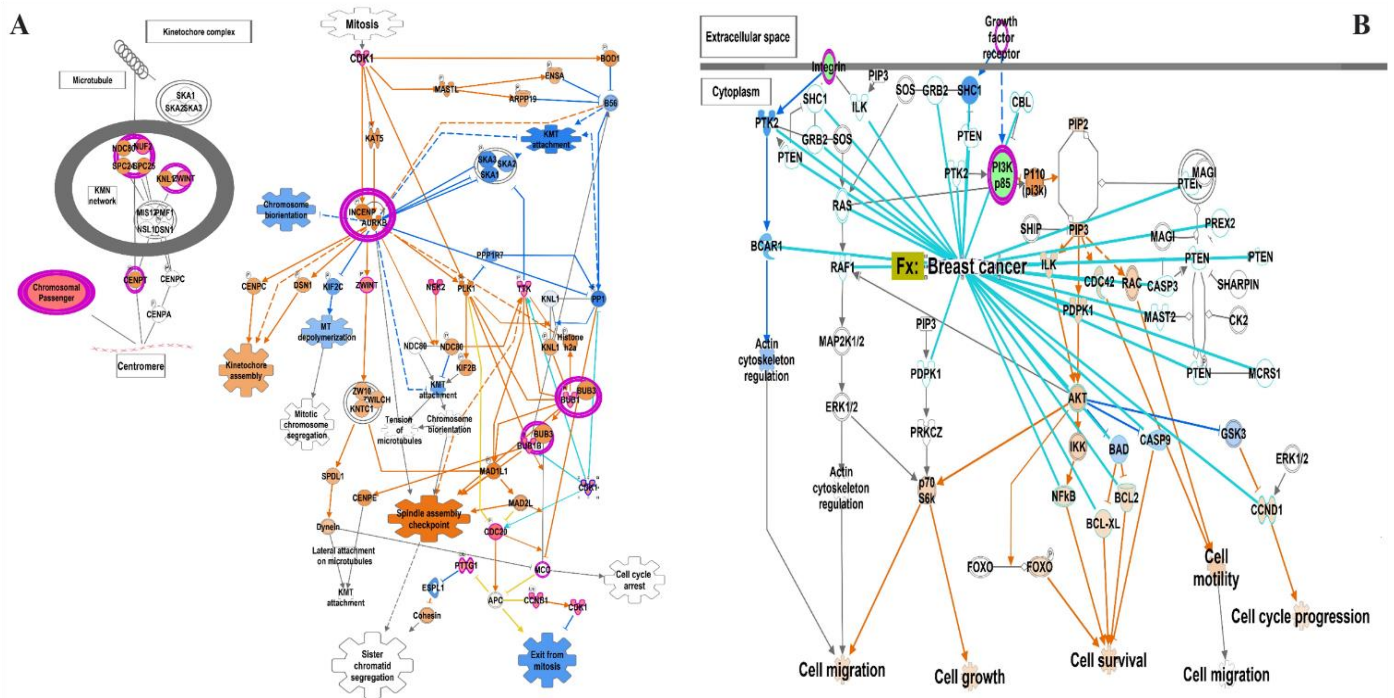


Figure 3. Canonical pathways derived using the IPA tool. (A) Kinetochore metaphase signaling pathway, (B) PTEN pathway overlapped with breast cancer associated genes.

Table 3. Gene ontology of biological processes derived from the enrichment of differentially expressed genes in breast cancer.

| Gene Set | Description | Gene Set Size | Expect Values | Overlap Value | Enrichment Ratio | FDR |
|------------|----------------------------------|---------------|---------------|---------------|------------------|------------------------|
| GO:0031012 | Extracellular matrix | 487 | 09.48 | 35 | 3.69 | 3.31×10^{-8} |
| GO:0051301 | Cell division | 576 | 11.21 | 41 | 3.65 | 2.69×10^{-9} |
| GO:1903047 | Mitotic cell cycle process | 780 | 15.18 | 46 | 3.02 | 2.62×10^{-8} |
| GO:0016477 | Cell migration | 1352 | 26.32 | 68 | 2.58 | 1.45×10^{-9} |
| GO:0042127 | Regulation of cell proliferation | 1535 | 29.88 | 72 | 2.40 | 3.75×10^{-9} |
| GO:0048870 | Cell motility | 1493 | 29.06 | 69 | 2.37 | 1.64×10^{-8} |
| GO:0051674 | Localization of cell | 1493 | 29.06 | 69 | 2.37 | 1.64×10^{-8} |
| GO:0009719 | Response to endogenous stimulus | 1574 | 30.64 | 71 | 2.31 | 2.01×10^{-8} |
| GO:0008283 | Cell proliferation | 1953 | 38.02 | 87 | 2.28 | 6.20×10^{-10} |
| GO:0009888 | Tissue development | 1814 | 35.31 | 77 | 2.18 | 3.31×10^{-8} |

3.3. Machine Learning Algorithms for the Identification of Diagnostic Biomarker Genes

Initially, seven ML algorithms predicted genes with diagnostic importance using 355 DEGs significantly expressed in BC samples, and important genes predicted by at least four ML models were selected for further analysis ($n = 65$). Additionally, we analyzed each dataset individually and checked the status of 355 DEGs in each dataset; the genes present in at least four datasets were selected for further analysis ($n = 94$). We identified 28 common genes passing both criteria (DEGs > 3 ML and DEGs > 4 datasets) as the potential hub of BC diagnostic and prognostic biomarkers (Figure 4A). With more stringent conditions (DEGs > 5 ML and DEGs > 7 dataset) and based on their role in tumorigenesis, a novel diagnostic nine-gene signature (*COL10A*, *S100P*, *ADAMTS5*, *WISP1*, *COMP*, *CXCL10*, *LYVE1*, *COL11A1*, and *INHBA*) was identified for BC. An unsupervised hierarchical clustering-based heatmap showed a correlation in a pairwise fashion between the samples and probes. A heatmap of unfiltered probes ($n = 54676$) was ambiguous and non-conclusive, while the unique set of DEGs ($n = 355$), hub genes ($n = 28$), and gene signatures ($n = 9$) for BC diagnosis showed a distinct correlation between the samples and gene expression (Figure 4B).

3.4. Machine-Learning-Algorithm-Based 10-Fold Cross-Validation

We used a 10-fold cross-validation technique to evaluate the performance of ML models for diagnostic and prognostic gene signatures. To perform 10-fold cross-validation, the dataset was divided into 10 equally sized folds. The model was then trained and validated 10 times, each time using a different fold for validation and the remaining nine folds for training. The process was repeated for all the folds, and the results were averaged to obtain an estimate of the model's performance. This provided a more reliable estimate of the model's performance compared to using a single-train test split, which may be biased based on the specific data that were selected; it can also help prevent overfitting.

For validating the diagnostic performance of our nine-gene signature, a new set of ML algorithms (GBDT, XGBoost, AdaBoost, KNN, and MLP) was employed to evaluate the model. By comparing the diagnostic efficiency, accuracy, and precision of different algorithms, the constructed diagnostic model was validated (Figure 5 and Table 4). Each ML model's performance was evaluated by measuring a range of performance metrics, including AUC, accuracy, precision, recall, and the F1 score. Here, all the ML methods predicted were above 95, and the ML model had the greatest AUC value indicating the candidate gene signature as a potential biomarker. KNN showed the highest values for all the evaluation metrics (mean F1 = 0.982), which indicates that it performed the best among

the five models. In contrast, MLP showed the lowest values for all evaluation metrics. Furthermore, biomarkers that could distinguish disease samples and normal samples were analyzed according to PCA, as it groups the samples based on similarities (Figure 6).

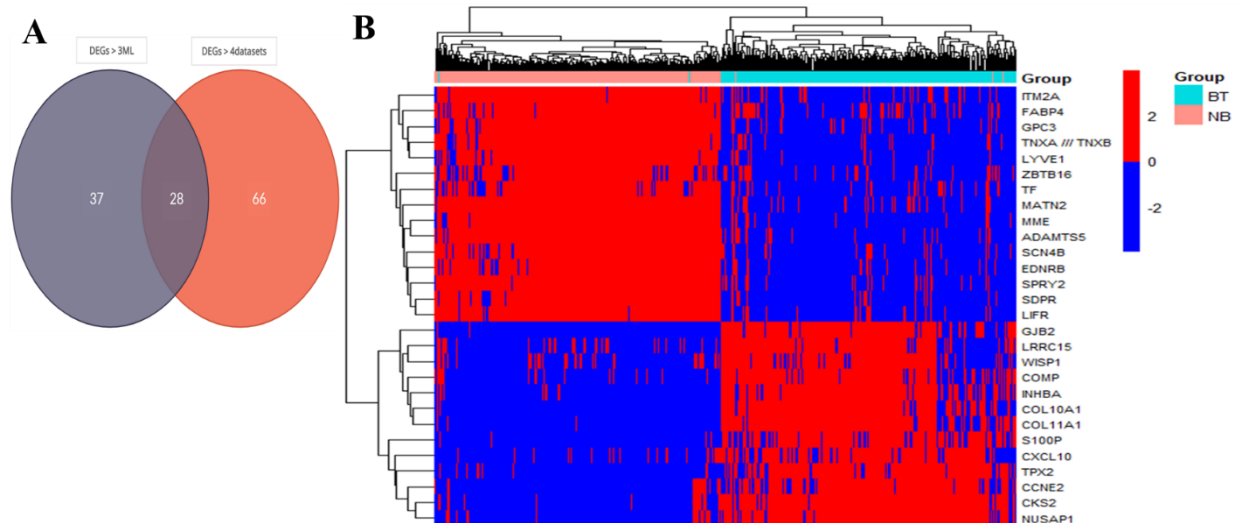


Figure 4. (A) Venn diagram showing 28 hub genes derived from the intersection of DEGs > 3 ML and DEGs > 4 datasets. (B) Unsupervised hierarchical clustering: heatmap of 701 samples, including 356 breast tumor (BT, cyan) and 345 normal breast (NB, pink) tissues, showing the gene expression pattern of 28 hub genes, including diagnostic and prognostic gene signatures. Upregulated genes are shown in red and downregulated genes are in blue.

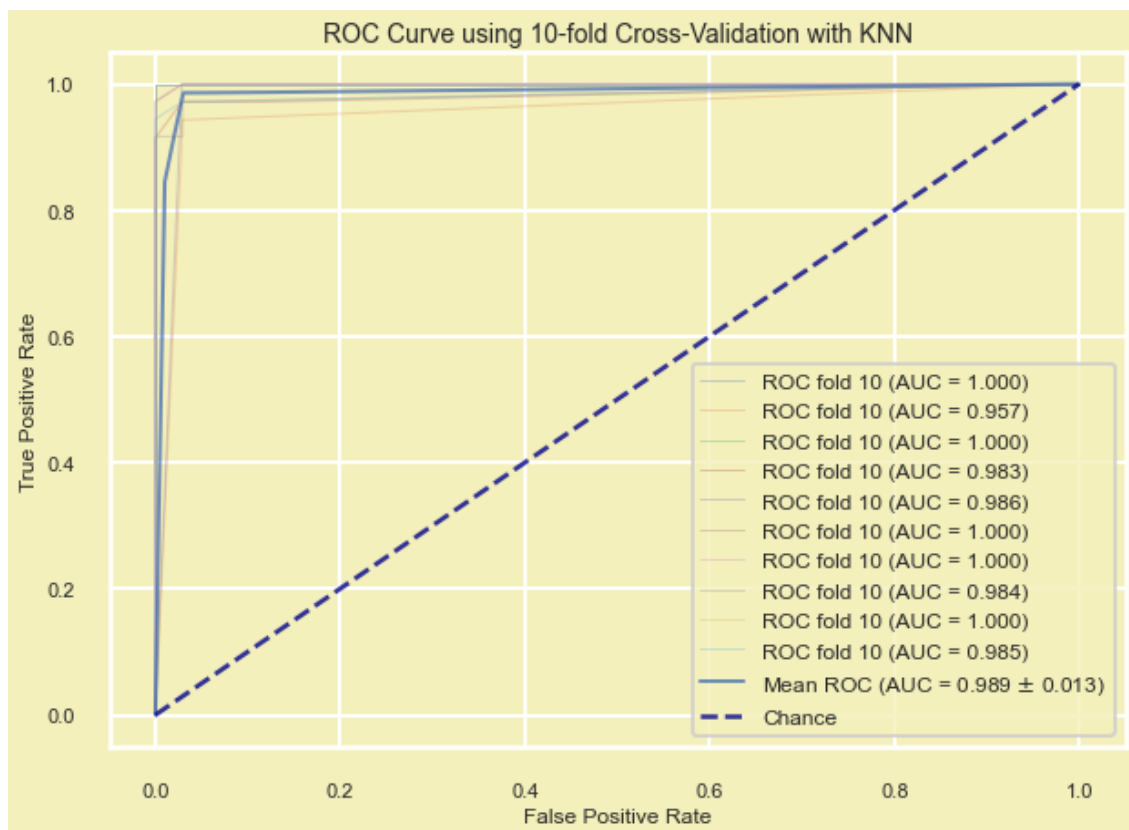
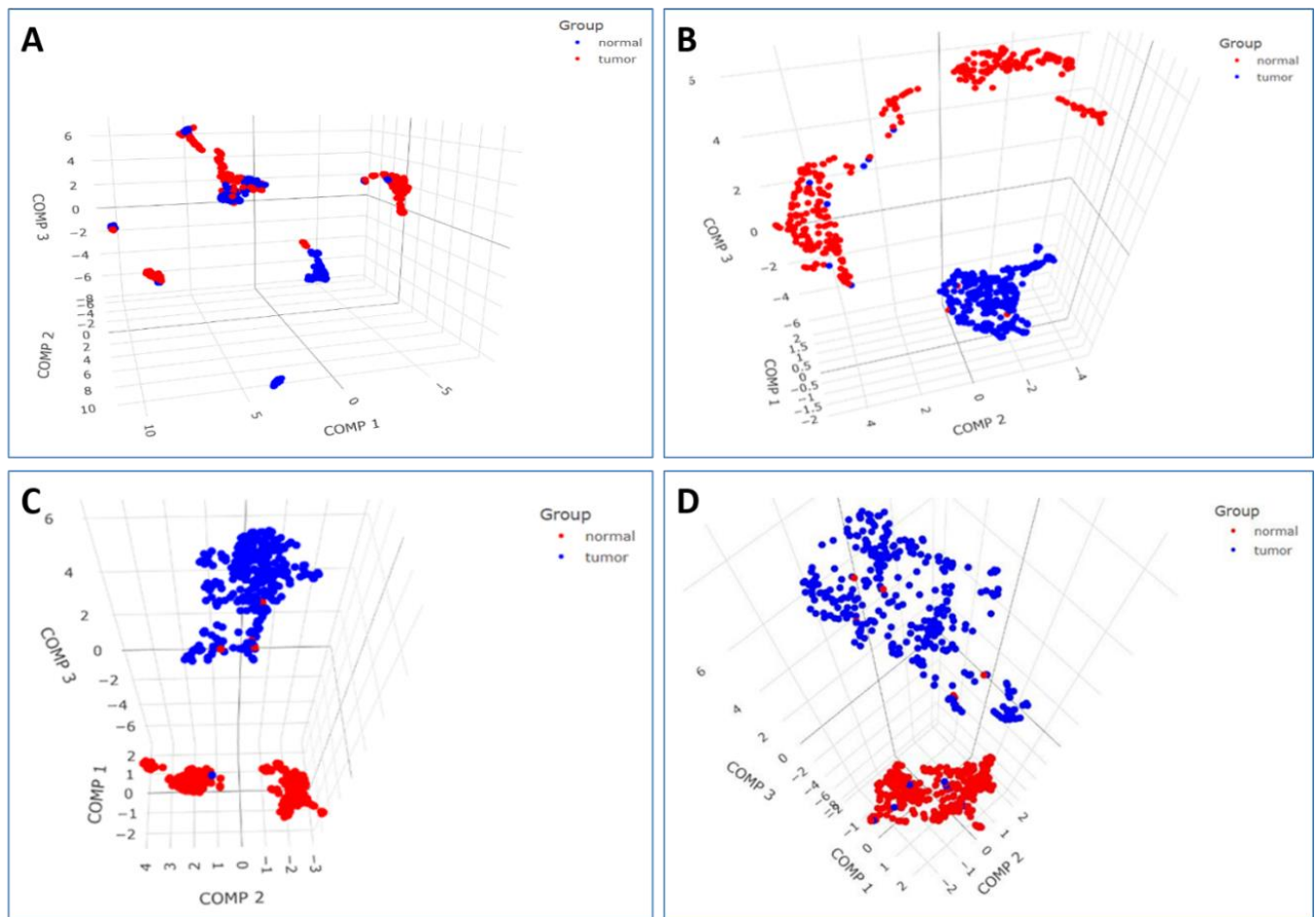


Figure 5. K-nearest neighbors (KNN)-based ML model for diagnostic gene signature showing the mean ROC (AUC 0.989 ± 0.013).

Table 4. Machine learning methods for the 10-fold cross-validation of the diagnostic nine-gene signature.

| ML Model | Mean AUC | Mean ACC | Mean Precision | Mean Recall | Mean F1 |
|----------|----------|----------|----------------|-------------|---------|
| KNN | 0.989 | 0.981 | 0.983 | 0.980 | 0.982 |
| GBDT | 0.995 | 0.973 | 0.970 | 0.978 | 0.973 |
| AdaBoost | 0.992 | 0.974 | 0.972 | 0.977 | 0.975 |
| XGBoost | 0.994 | 0.971 | 0.969 | 0.975 | 0.972 |
| MLP | 0.975 | 0.960 | 0.961 | 0.961 | 0.960 |

**Figure 6.** PCA plot showing an overall distribution of the samples ($n = 701$), including breast tumor (blue) and normal breast tissue (red) based on transcriptomics profiles: (A) 54,675 probes, (B) 355 DEGs, (C) 28 hub genes, and (D) diagnostic nine-gene signature.

3.5. Survival Analysis to Identify Genes with Prognostic Importance

Survival analysis using the KM estimator was conducted for 28 hub genes. First, relapse-free survival and overall survival analyses were performed for mRNA (gene-chip), followed by the overall survival for mRNA (RNA seq). We identified a novel prognostic gene signature of eight genes (*CCNE2*, *NUSAP1*, *TPX2*, *S100P*, *ITM2A*, *LIFR*, *TNXA*, and *ZBTB16*) that were significant (log-rank p -value < 0.05) under both RFS and OS conditions (Tables 5 and 6, and Figures 7 and 8). The hazard ratio (HR) compares the risk of death (overall survival) and postoperative follow-up (relapse-free survival) occurring between the high and low expressions of the gene. The HR value < 1 or > 1 indicates that the risks associated with a lower expression and higher expression of the gene are significantly different. On the other hand, the confidence intervals (CI) indicate the level of uncertainty around the estimated survival probability at each time point. A narrower confidence

interval indicates that the survival estimate is more precise and that the sample size is large enough to produce reliable results. The CI value demonstrates the precision and reliability of the results. Finally, the log-rank p -value measures the statistical significance that helps determine whether the observed difference in survival between the groups (high and low expression of the gene) is statistically significant (<0.05), i.e., unlikely to have occurred by chance. Therefore, log-rank p -values are the deciding factors for survival significance.

Table 5. mRNA (gene chip) and the relapse-free survival analysis of 28 hub genes, with the measured hazard ratio (HR), confidence interval (CI), and log-rank p -value.

| Gene Symbol | Probe_IDs | HR | CI | Log-Rank p -Value | Decision (Log-Rank p -Value) |
|----------------|-------------|-----|-----------|------------------------|--------------------------------|
| <i>ADAMTS5</i> | 219935_at | 0.9 | 0.77–0.94 | 1.50×10^{-3} | Significant |
| <i>CCNE2</i> | 205034_at | 1.9 | 1.67–2.06 | 1.00×10^{-16} | Significant |
| <i>CKS2</i> | 204170_s_at | 1.7 | 1.51–1.85 | 1.00×10^{-16} | Significant |
| <i>CXCL10</i> | 204533_at | 1.2 | 1.12–1.37 | 4.40×10^{-5} | Significant |
| <i>EDNRB</i> | 206701_x_at | 0.8 | 0.69–0.85 | 2.20×10^{-7} | Significant |
| <i>FABP4</i> | 203980_at | 0.9 | 0.81–0.99 | 2.58×10^{-2} | Significant |
| <i>GPC3</i> | 209220_at | 0.8 | 0.76–0.92 | 5.00×10^{-4} | Significant |
| <i>ITM2A</i> | 202747_s_at | 0.7 | 0.63–0.77 | 1.40×10^{-12} | Significant |
| <i>LIFR</i> | 225575_at | 0.7 | 0.56–0.75 | 1.60×10^{-8} | Significant |
| <i>MATN2</i> | 202350_s_at | 0.9 | 0.78–0.95 | 3.30×10^{-3} | Significant |
| <i>LYVE1</i> | 219059_s_at | 0.9 | 0.81–0.99 | 3.78×10^{-2} | Significant |
| <i>NUSAP1</i> | 218039_at | 1.7 | 1.54–1.89 | 1.00×10^{-16} | Significant |
| <i>SCN4B</i> | 236359_at | 0.6 | 0.55–0.75 | 1.00×10^{-8} | Significant |
| <i>SDPR</i> | 222717_at | 0.7 | 0.57–0.77 | 9.70×10^{-8} | Significant |
| <i>SPRY2</i> | 204011_at | 0.9 | 0.79–0.97 | 1.02×10^{-2} | Significant |
| <i>TF</i> | 214063_s_at | 0.9 | 0.78–0.96 | 5.20×10^{-3} | Significant |
| <i>TNXA</i> | 216333_x_at | 0.7 | 0.63–0.77 | 2.40×10^{-12} | Significant |
| <i>TPX2</i> | 210052_s_at | 1.6 | 1.48–1.82 | 1.00×10^{-16} | Significant |
| <i>WISP1</i> | 229802_at | 0.8 | 0.64–0.87 | 1.00×10^{-4} | Significant |
| <i>ZBTB16</i> | 205883_at | 0.7 | 0.58–0.72 | 1.00×10^{-16} | Significant |
| <i>COL11A1</i> | 37892_at | 1.2 | 1.12–1.38 | 2.30×10^{-5} | Significant |
| <i>INHBA</i> | 210511_s_at | 1.2 | 1.06–1.3 | 1.70×10^{-3} | Significant |
| <i>S100P</i> | 204351_at | 1.5 | 1.31–1.61 | 6.30×10^{-3} | Significant |
| <i>COL10A1</i> | 205941_s_at | 1.0 | 0.88–1.08 | 6.60×10^{-1} | Insignificant |
| <i>COMP</i> | 205713_s_at | 0.9 | 0.85–1.04 | 2.53×10^{-1} | Insignificant |
| <i>GJB2</i> | 223278_at | 1.0 | 0.88–1.19 | 7.89×10^{-1} | Insignificant |
| <i>LRRC15</i> | 213909_at | 0.9 | 0.82–1.01 | 7.16×10^{-2} | Insignificant |
| <i>MME</i> | 203435_s_at | 1.1 | 0.98–1.2 | 1.28×10^{-1} | Insignificant |

Table 6. mRNA (gene chip) and the overall survival analysis of 28 hub genes, with the measured hazard ratio (HR), confidence interval (CI), and log-rank *p*-value.

| Gene Symbol | Probe_IDs | HR | CI | Log-Rank <i>p</i> -Value | Decision (Log-Rank <i>p</i> -Value) |
|----------------|-------------|------|-----------|--------------------------|-------------------------------------|
| <i>CCNE2</i> | 205034_at | 1.47 | 1.22–1.78 | 5.00×10^{-5} | Significant |
| <i>CKS2</i> | 204170_s_at | 1.32 | 1.09–1.59 | 3.70×10^{-3} | Significant |
| <i>ITM2A</i> | 202747_s_at | 0.61 | 0.5–0.73 | 2.00×10^{-7} | Significant |
| <i>LIFR</i> | 225575_at | 0.59 | 0.45–0.78 | 1.00×10^{-4} | Significant |
| <i>NUSAP1</i> | 218039_at | 1.65 | 1.36–2 | 1.90×10^{-7} | Significant |
| <i>SDPR</i> | 222717_at | 0.70 | 0.53–0.92 | 8.90×10^{-3} | Significant |
| <i>TNXA</i> | 216333_x_at | 0.71 | 0.59–0.85 | 3.00×10^{-4} | Significant |
| <i>TPX2</i> | 210052_s_at | 1.56 | 1.29–1.89 | 3.20×10^{-6} | Significant |
| <i>ZBTB16</i> | 205883_at | 0.63 | 0.52–0.76 | 1.40×10^{-6} | Significant |
| <i>S100P</i> | 204351_at | 1.50 | 1.25–1.82 | 2.10×10^{-5} | Significant |
| <i>ADAMTS5</i> | 219935_at | 0.85 | 0.7–1.02 | 8.00×10^{-2} | Insignificant |
| <i>COL10A1</i> | 205941_s_at | 0.96 | 0.79–1.15 | 6.38×10^{-1} | Insignificant |
| <i>COMP</i> | 205713_s_at | 1.06 | 0.88–1.27 | 5.67×10^{-1} | Insignificant |
| <i>CXCL10</i> | 204533_at | 0.9 | 0.75–1.09 | 2.98×10^{-1} | Insignificant |
| <i>EDNRB</i> | 206701_x_at | 0.88 | 0.73–1.06 | 1.85×10^{-1} | Insignificant |
| <i>FABP4</i> | 203980_at | 0.84 | 0.7–1.02 | 7.78×10^{-2} | Insignificant |
| <i>GJB2</i> | 223278_at | 1.18 | 0.9–1.54 | 2.29×10^{-1} | Insignificant |
| <i>GPC3</i> | 209220_at | 0.84 | 0.7–1.02 | 7.47×10^{-2} | Insignificant |
| <i>MATN2</i> | 202350_s_at | 0.85 | 0.7–1.02 | 8.10×10^{-2} | Insignificant |
| <i>LRRC15</i> | 213909_at | 0.87 | 0.72–1.04 | 1.30×10^{-1} | Insignificant |
| <i>LYVE1</i> | 219059_s_at | 1.04 | 0.86–1.25 | 6.90×10^{-1} | Insignificant |
| <i>MME</i> | 203435_s_at | 0.83 | 0.69–1.01 | 5.95×10^{-2} | Insignificant |
| <i>SCN4B</i> | 236359_at | 0.83 | 0.63–1.08 | 1.68×10^{-1} | Insignificant |
| <i>SPRY2</i> | 204011_at | 0.89 | 0.74–1.08 | 2.36×10^{-1} | Insignificant |
| <i>TF</i> | 214063_s_at | 0.99 | 0.82–1.19 | 9.08×10^{-1} | Insignificant |
| <i>WISP1</i> | 229802_at | 0.79 | 0.6–1.03 | 8.33×10^{-2} | Insignificant |
| <i>COL11A1</i> | 37892_at | 1.12 | 0.93–1.35 | 2.37×10^{-1} | Insignificant |
| <i>INHBA</i> | 210511_s_at | 1.12 | 0.93–1.36 | 0.2212 | Insignificant |
| <i>CKS2</i> | 204170_s_at | 1.32 | 1.09–1.59 | 0.0612 | Insignificant |
| <i>SDPR</i> | 222717_at | 0.7 | 0.53–0.92 | 0.0628 | Insignificant |

First, we validated the prognostic eight-gene signature using mRNA (RNA seq) dataset based on the RFS and OS analyses by collectively measuring the hazard ratio (HR), confidence interval (CI), and log-rank *p*-value (Figure 9 and Table 7), and the prognostic signatures were again validated using five ML methods, including GBDT, XGBoost, AdaBoost, KNN, and MLP. Among the five ML models, GBDT showed the highest values for the mean AUC (0.993), mean accuracy (0.980), and mean F1 score (0.98), while XGBoost

showed the highest mean precision (0.981). KNN showed the second-highest values for all the evaluation metrics, while MLP showed the lowest values (Figure 10 and Table 8).

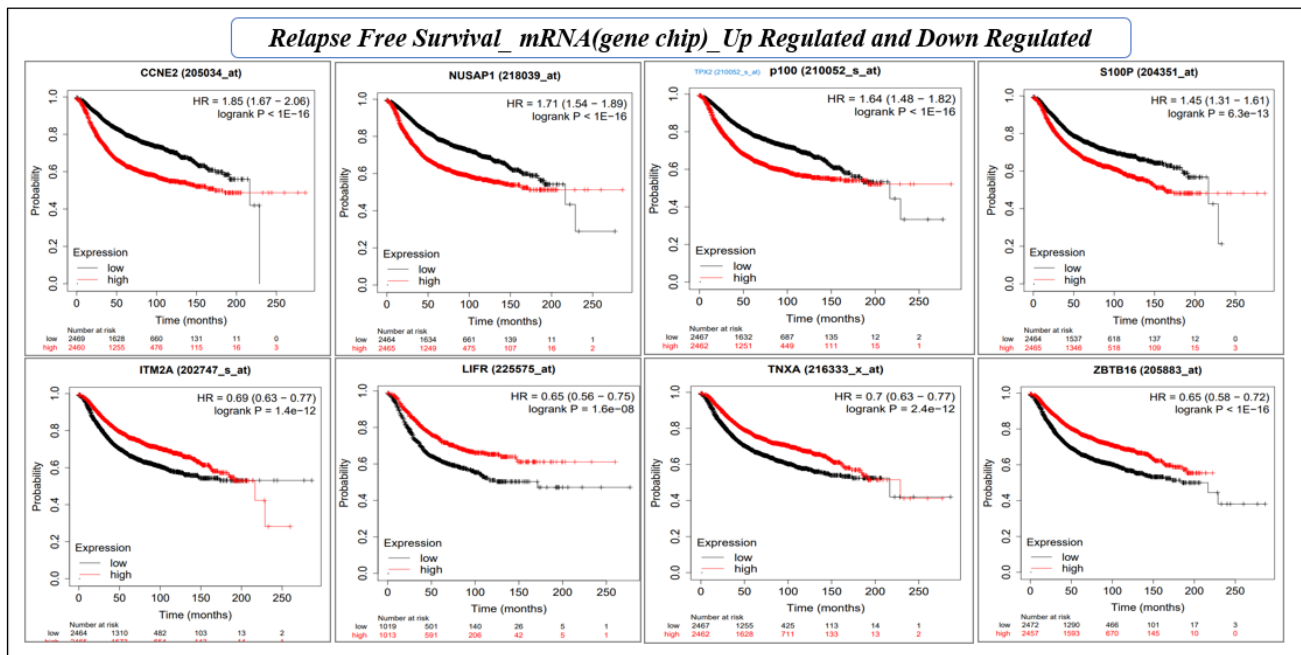


Figure 7. KM plot based on the relapse-free survival analysis of eight individual genes (mRNA, gene-chip) of prognostic gene signature. The X-axis and Y-axis represent time in months and the probability of the survival of patients, respectively. The impact of the high and low expression of the gene on patient survival is shown in red and black lines, respectively.

Table 7. mRNA (RNA seq) based on the relapse-free survival and overall survival analyses of eight prognostic gene hubs collectively measuring the hazard ratio (HR), confidence interval (CI), and log-rank *p*-value.

| Survival Type | Gene Hub | HR | CI | Log-Rank <i>p</i> -Value | Decision by Log-Rank <i>p</i> -Value | Expression |
|---------------|---------------------------------------|------|-----------|--------------------------|--------------------------------------|---------------|
| RFS | <i>CCNE2, NUSAP1, TPX2, S100P</i> | 1.62 | 1.46–1.79 | 1.00×10^{-16} | Significant | Upregulated |
| RFS | <i>ITM2A, LIFR, TNXA-TNXB, ZBTB16</i> | 0.58 | 0.50–0.68 | 1.90×10^{-12} | Significant | Downregulated |
| OS | <i>CCNE2, NUSAP1, TPX2, S100P</i> | 1.44 | 1.19–1.74 | 0.00014 | Significant | Upregulated |
| OS | <i>ITM2A, LIFR, TNXA-TNXB, ZBTB16</i> | 0.57 | 0.43–0.75 | 4.20×10^{-5} | Significant | Downregulated |

Table 8. Machine learning model for the 10-fold cross-validation of the prognostic eight-gene signature.

| ML Model | Mean_AUC | Mean_ACC | Mean_Precision | Mean_Recall | Mean_F1 |
|----------|----------|----------|----------------|-------------|---------|
| GBDT | 0.993 | 0.980 | 0.983 | 0.977 | 0.980 |
| XGBoost | 0.992 | 0.976 | 0.981 | 0.972 | 0.976 |
| AdaBoost | 0.987 | 0.967 | 0.965 | 0.972 | 0.968 |
| KNN | 0.985 | 0.979 | 0.978 | 0.980 | 0.979 |
| MLP | 0.979 | 0.966 | 0.975 | 0.958 | 0.966 |

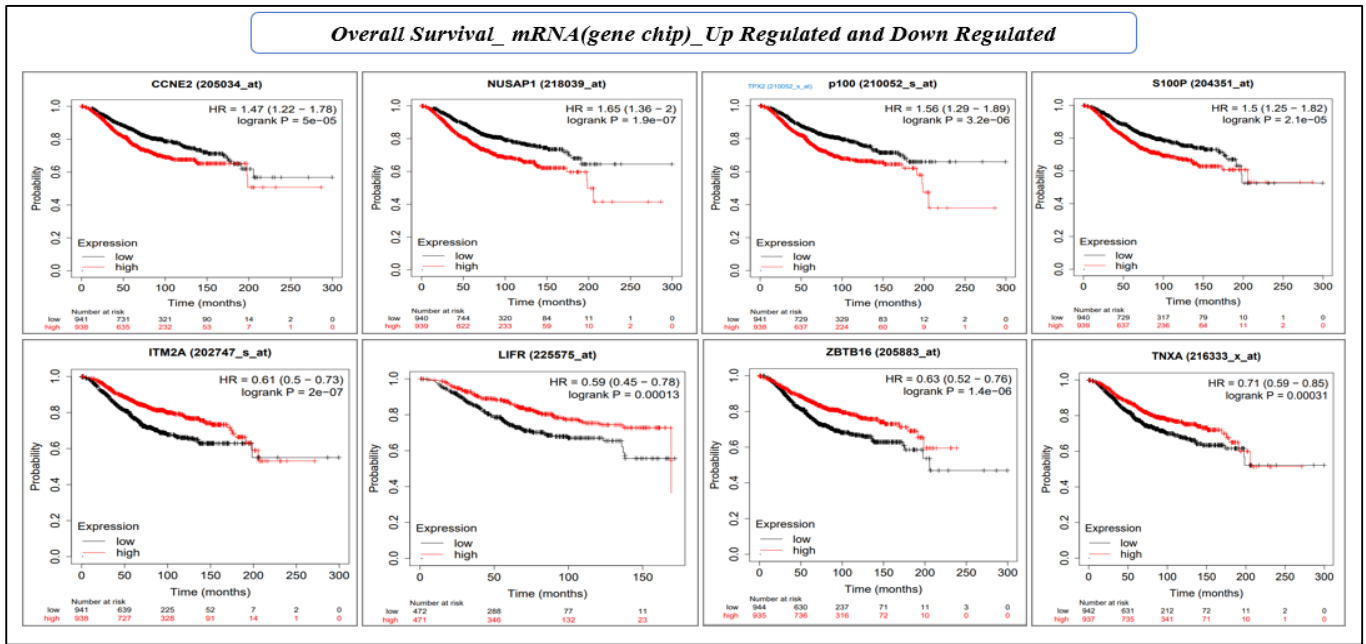


Figure 8. KM plot based on the overall survival analysis of eight individual genes (mRNA, gene-chip) of prognostic gene signature. The X-axis and Y-axis represent time in months and the probability of the survival of patients, respectively. The impact of the high and low expression of the gene on patient survival is shown in red and black lines, respectively.

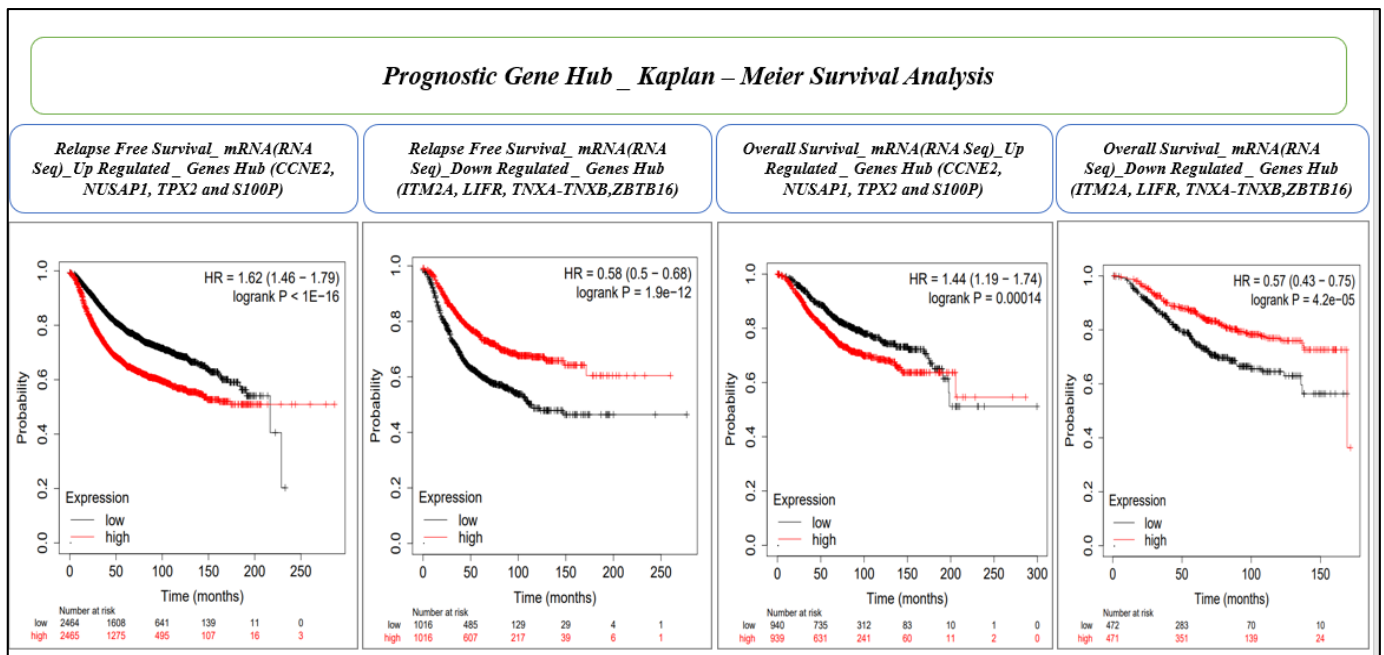


Figure 9. RFS and OS analyses and the validation of upregulated (CCNE2, NUSAP1, TPX2, and S100P), and downregulated (ITM2A, LIFR, TNXA, and ZBTB16) gene groups (mRNA, RNA seq) of the prognostic gene signature. The X-axis and Y-axis represent time in months and the probability of the survival of patients. The impact of the high and low expression of the gene on patient survival is shown in red and black lines, respectively.

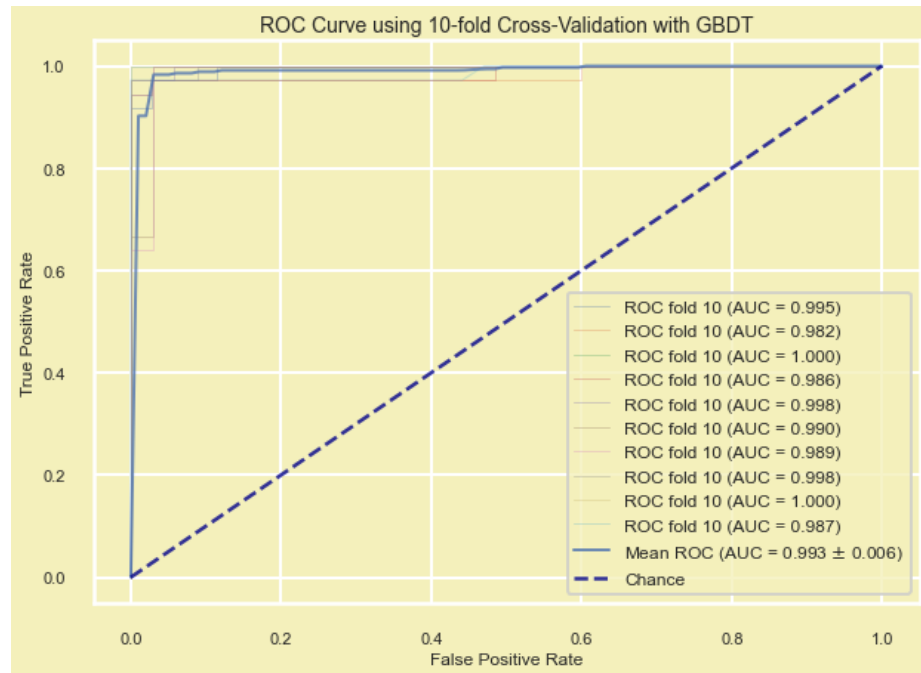


Figure 10. Gradient-boosting decision trees (GBDT) based on the ML model for the prognostic gene signature showing the mean ROC (AUC 0.993 ± 0.006).

3.6. qRT-PCR Analysis

qPCR was used to confirm the identified BC biomarkers (gene signatures) by determining the relative expression of 16 genes (nine-gene signature for diagnosis: *COL10A*, *S100P*, *ADAMTS5*, *WISP1*, *COMP*, *CXCL10*, *LYVE1*, *COL11A1*, and *INHBA*; and eight-gene signature for prognosis: *CCNE2*, *NUSAP1*, *TPX2*, *S100P*, *ITM2A*, *LIFR*, *TNXA*, and *ZBTB16*), with an overlap of the *S100P* gene (Figure 11). *GAPDH* was used as the internal control. We found qPCR results in concordance with microarray-analyzed expression patterns (Table 9).

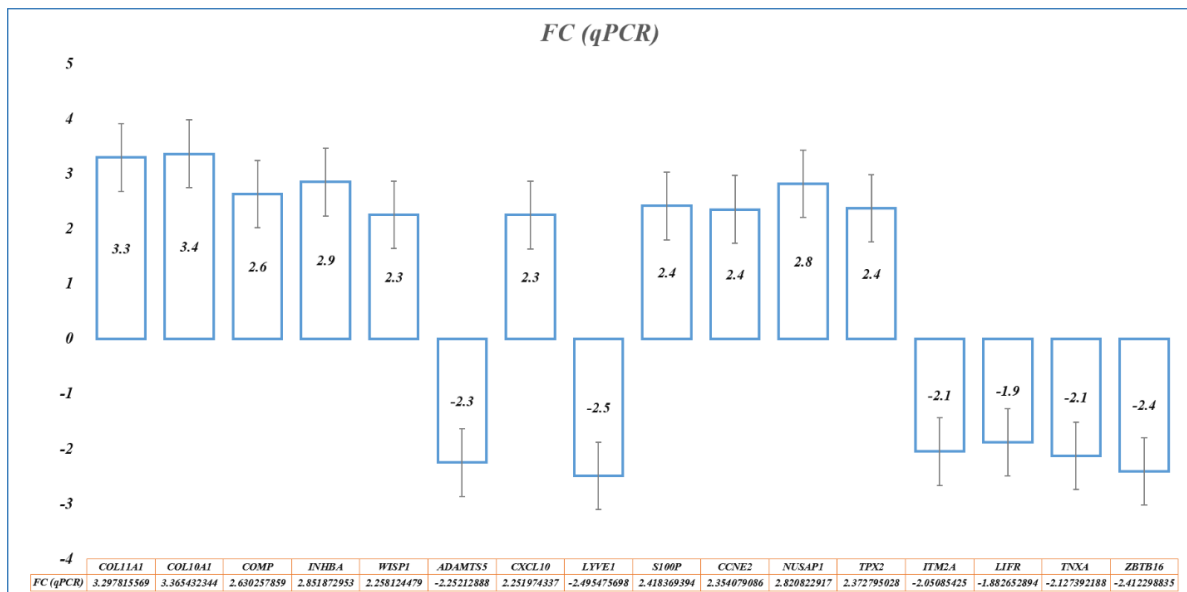


Figure 11. qRT-PCR results showing overexpression of *COL10A*, *S100P*, *WISP1*, *COMP*, *CXCL10*, *COL11A1*, *INHBA*; *CCNE2*, *NUSAP1*, *TPX2*, and *S100P* genes, and under-expression of *ADAMTS5*, *LYVE1*, *ITM2A*, *LIFR*, *TNXA*, and *ZBTB16* genes.

Table 9. Expression of gene signatures in microarray and qRT-PCR. The results of qRT-PCR are represented by the quantitative expression (Rq) and fold-change (FC), along with the standard deviation (StdDev) and *p*-values.

| Gene | Microarray | | qRT-PCR | | | |
|--|------------|-------------------------|----------|----------|----------|------------------------|
| | FC | adj. <i>p</i> -Value | Rq | FC | StdDev | <i>p</i> -Value |
| Expression of Diagnostic Gene Signature | | | | | | |
| COL11A1 | 3.907519 | 5.1×10^{-163} | 9.834254 | 3.297816 | 0.64495 | 3.47×10^{-7} |
| COL10A1 | 3.842333 | 2.1×10^{-178} | 10.30614 | 3.365432 | 0.660567 | 1.84×10^{-8} |
| S100P | 3.701498 | 3.6×10^{-137} | 5.345665 | 2.418369 | 0.995349 | 2.73×10^{-11} |
| COMP | 3.150415 | 6.3×10^{-137} | 6.191366 | 2.630258 | 0.425433 | 5.81×10^{-15} |
| INHBA | 3.042628 | 2.6×10^{-157} | 7.21937 | 2.851873 | 0.901961 | 1.97×10^{-8} |
| WISP1 | 2.551547 | 6×10^{-105} | 4.783692 | 2.258124 | 0.461439 | 2.9×10^{-8} |
| ADAMTS5 | −3.13169 | 3.3×10^{-184} | 0.209914 | −2.25213 | 0.511629 | 2.88×10^{-9} |
| CXCL10 | 2.530934 | 2.16×10^{-95} | 4.763343 | 2.251974 | 0.86944 | 2.16×10^{-5} |
| LYVE1 | −3.14204 | 1.2×10^{-142} | 0.177332 | −2.49548 | 0.877592 | 1.73×10^{-6} |
| Expression of Prognostic Gene Signature | | | | | | |
| CCNE2 | 2.530327 | 3.7×10^{-154} | 5.112678 | 2.354079 | 0.392123 | 3.1×10^{-15} |
| NUSAP1 | 2.732299 | 2.4×10^{-124} | 7.065653 | 2.820823 | 0.9127 | 3.81×10^{-10} |
| TPX2 | 2.145025 | 5.6×10^{-135} | 5.179436 | 2.372795 | 0.432888 | 4.72×10^{-10} |
| ITM2A | −2.54576 | 9.1×10^{-149} | 0.241341 | −2.05085 | 0.683145 | 4.21×10^{-8} |
| LIFR | −3.0494 | 5.6×10^{-159} | 0.271185 | −1.88265 | 0.853309 | 1.78×10^{-6} |
| TNXA | −2.54523 | 1.9×10^{-129} | 0.228871 | −2.12739 | 0.736176 | 7.89×10^{-7} |
| ZBTB16 | −2.4943 | 1.12×10^{-115} | 0.187856 | −2.4123 | 0.567998 | 3.32×10^{-9} |
| S100P | 3.701498 | 3.6×10^{-137} | 5.345665 | 2.418369 | 0.995349 | 2.73×10^{-11} |

4. Discussion

In recent years, multiple molecular diagnostic prognostic and predictive biomarkers have been proposed, and despite the availability of few molecular tests, traditional pathological factors such as the number of lymph node metastases, tumor size, and tumor grade, continue to be mandatory for clinical decisions [38]. However, in the era of personalized treatment, these factors alone are inadequate and require molecular/genomic assistance, as cancer occurs via genetic alterations that transform normal cells into tumor cells. Although significant knowledge exists related to carcinogenesis, a complete understanding of cancer development mechanisms is still required. In recent years, genomics and proteomics have played a vital part in the development of different biomarkers for breast cancer [39,40]. Gene expression profiling can detect genetic alterations in the origin, growth, proliferation, and metastasis of tumors, and classify them accordingly. Gene expression signatures, derived from DEGs, specifically correlate these genetic alterations with clinical variables such as the diagnosis and prognosis [20,41]. A correct and timely diagnosis is the starting point of treatment and determination of prognosis is the most immediate challenge in patient management. This can be best achieved through a combination of traditional clinicopathological prognostic factors, molecular biomarkers such as single-gene tests (ER, PR, HER2) and specific multigene tests (gene signatures).

Among the 355 DEGs identified in a combined BC cohort, *COL11A1*, *TOP2A*, *S100P*, *COL10A1*, and *RRM2* were the most upregulated, while *ADH1B*, *ADIPOQ*, *PLIN1*, *LEP*, and *LPL* were the most downregulated DEGs in BC. The pathway and enrichment analyses of DEGs revealed activation of the kinetochore metaphase signaling pathway, PTEN path-

way, HOTAIR regulatory pathway, etc., and suppression of the senescence pathway and phagosome formation pathways in BC. The most significantly enriched molecular processes were the extracellular matrix, cell division, mitotic cell cycle process, cell migration, and regulation of cell proliferation. First, to verify the reliability of our method of screening for biomarkers, we confirmed our finding of DEGs, pathways, and gene ontologies using literature mining and verification. Matching our results with previous findings was good evidence that they are indeed involved in the development and progression of BC [42–46].

Kinetochores architecture and its functional regulation is one of the most fascinating multi-protein machineries in a cell [47]. The kinetochore metaphase signaling is essential for chromosome segregation in mitosis and meiosis [48]. The critical regulators of alignment and segregation of chromosomes during mitosis, aurora B kinase (AURKB), dual specificity protein kinase TTK (Mps1), and kinetochore protein NDC80 homolog (NDC80) previously reported were significant in our study too [49]. Another essential pathway that was significantly upregulated in our study was PTEN/PI3K/AKT. This controls the signaling of numerous biological processes, including apoptosis, cell proliferation, cell growth, and metabolism. Phosphatase and tensin homolog deleted on chromosome 10 (PTEN) is a dual protein/lipid phosphatase, of which the main substrate is phosphatidylinositol(3,4,5) triphosphate (PIP3), the product of PI3K [50,51]. The PTEN tumor suppressor is the chief brake of the PI3K-Akt pathway and a common target for inactivation in somatic cancers [52]. PTEN activity is frequently lost in several metastatic human cancers due to mutations, deletions, or promoter methylation silencing [50]. Senescence is associated with mitochondrial metabolic activities such as the tricarboxylic acid cycle, oxidative phosphorylation, and glycolytic pathways. The old senescent cells die during aging or apoptosis. The senescence pathway promotes cell cycle arrest triggered in response to stress with increased AMP/ADP:ATP and NAD⁺/NADH ratios, and activating AMPK, p53, p16, KRAS, etc. [53–55]. The *in vitro* demonstration of oncogene-induced senescence establishes senescence as a vital tumor-suppressive mechanism, in addition to apoptosis. Senescence not only stops the proliferation of premalignant cells (tumorigenesis), but also eases the clearance of affected cells through the immunosurveillance [56]. *In vivo* studies showed that suppression of the senescence pathway can also promote mammary tumorigenesis [57].

AI and ML techniques based on automated medical diagnosis are increasing gradually for clinical, pathological, and radiological reports. The fusion of multiple techniques in different types of data processing for cancer study must be a further instrument to obtain successful results. An earlier convolution neural network approach had been applied for image processing in medical diagnosis [58]. However, using AI and ML in the evaluation of high-throughput genomics data from patients in diagnostic decision-making is still a bottle neck in healthcare [59–61]. Typically, microarray data have thousands of features (genes/probes), but only a few samples (in tens or hundreds). For ML classification, it is better to have a large cohort with fewer features. Eleven BC datasets from different studies were integrated to increase the cohort size. Transcriptomics profiling resulted in 355 DEGs associated with BC, but this number was technically too big to recommend for gene signature biomarkers for diagnostic or prognostic tests. However, AI and ML have the potential to filter out genes with the best diagnostic and prognostic importance. Thus, for BC diagnosis via binary classification (whether or not BC), we used seven ML and feature selection methods (RFECV-LR, RFECV-SVM, RF, extra trees, LASSO, SVM-L1, SVM-L2, and GA) for gene reduction, and found high accuracy in the models. We identified a hub of 28 genes predicted by at least three ML methods and present in at least four BC datasets. RFECV-LR and RFECV-SVM improved the classification accuracy of logistic regression by selecting the most relevant features for the model and helped reduce overfitting by removing irrelevant or redundant features. Recursive feature elimination was utilized to rank the genes, with a random forest classifier used to evaluate gene fitness through five-fold cross-validation [62]. The extra trees method was versatile, less prone to overfitting, computationally efficient, robust to noise, and could handle missing data [29,63]. LASSO had advantages in its ability to handle multicollinearity, and provided a sparse solution for

both variable selection and shrinkage problems [64,65]. SVM models were frequently used for classification and regression tasks using L1 and L2 regularization [66]. L1 regularization had improved the interpretability of the model and reduced overfitting by encouraging sparsity and selecting only the most relevant features for the classification task. GA utilized the concept of survival of the fittest and was based on a population-based search approach for a robust and efficient search [67].

Based on the importance of the 28 hub genes in BC and using stringent filter conditions such as the genes predicted to be diagnostically important by at least five ML methods and present in at least seven BC datasets, a novel nine-gene signature (*COL10A*, *S100P*, *ADAMTS5*, *WISP1*, *COMP*, *CXCL10*, *LYVE1*, *COL11A1*, and *INHBA*) was identified. Similarly, by evaluating 28 hub genes using RFS and OS analyses by KM plot, a novel prognostic model consisting of an eight-gene signature (*CCNE2*, *NUSAP1*, *TPX2*, *S100P*, *ITM2A*, *LIFR*, *TNXA*, and *ZBTB16*) was identified. Many gene expression signatures have been proposed for BC diagnosis and prognosis in recent years; few are under trial and five of them succeeded to get FDA approval for commercial and clinical application, including OncotypeDX (21-gene signature), MammaPrint (70-gene signature), Prosigna (58-gene signature), EndoPredict (12-gene signature), and Breast Cancer Index (7-gene signature) [68,69].

Gene signature validation was crucial before recommendation for further analysis and clinical trials. We used 10-fold-cross validation by five ML methods including KNN, GBDT, AdaBoost, XGBoost, and MLP. Several studies have reported successful applications of these ML methods in BC gene hub/signature validation [70–75]. KNN-based validation was used to classify genes based on their expression profiles, and identify the gene clusters associated with cancer metastasis [72,73]. GBDT and AdaBoost were used to identify the key genes and pathways associated with breast cancer metastasis [72,74,76]. In addition, to identify disease-associated genes and pathways, XGBoost predicted cancer recurrence based on gene expression data [71,73]. The MLP method predicted gene clusters from expression data that were functionally related and associated with BC [70,72].

CCNE2 (Cyclin E2), involved in cell cycle regulation, can serve as an individual indicator of the likely outcome for BC patients. It is upregulated in tumor tissues and has the potential to function as a biomarker and linked to worse metastasis-free survival (MFS) outcomes and a poor overall survival [77,78]. *NUSAP1* (nucleolar and spindle-associated protein 1) playing a critical role in cell division and being a useful prognostic marker, has been implicated in various types of cancer, including BC [79]. The *TPX2* (targeting protein for xenopus kinesin-like protein 2), a microtubule-associated protein involved in spindle formation and cell division, is highly expressed in various cancers, including BC [80]. A high expression of *TPX2* can reduce the survival time of HER2-positive patients, as well as triple negative BC [81]. *S100P*, a calcium-binding protein, is involved in cell proliferation, differentiation, and apoptosis. An overexpression of *S100P* in BC cells makes it more aggressive, and hence it has the potential as a prognostic and therapeutic biomarker [8]. *ITM2A* (integral membrane protein 2A) regulates cellular growth and survival, and its low expression may play a role in the progression of BC, especially at advanced stages and higher grades of triple-negative breast cancer [82]. A decreased expression of *LIFR* (leukemia inhibitory factor receptor) may be a marker for a poorer prognosis and reduced survival in BC [83]. *TNXA* (tenascin XA) is an extracellular matrix protein involved in cell adhesion and migration. The survival analysis of abnormally expressed *TNXA* in breast tissue indicates poor prognosis [84]. A low expression of *ZBTB16* (zinc finger and BTB domain-containing protein 16) in BC has been associated with a poor prognosis and an increased risk of metastasis [85,86].

5. Conclusions

Artificial intelligence and machine learning approaches for the identification of novel diagnostic and prognostic gene signature biomarkers for breast cancer using microarray-based gene expression profiles were attempted. Initially, we identified a total of 355 DEGs

via gene expression profiling of BC microarray data, and our artificial-intelligence-based strategy significantly reduced the number of genes needed for an effective evaluation of diagnostic and prognostic importance. As a result, two novel gene signatures were highlighted, (i) diagnostic nine-gene signature (*COL10A*, *S100P*, *ADAMTS5*, *WISP1*, *COMP*, *CXCL10*, *LYVE1*, *COL11A1*, and *INHBA*) and (ii) prognostic eight-gene signature (*CCNE2*, *NUSAP1*, *TPX2*, *S100P*, *ITM2A*, *LIFR*, *TNXA*, and *ZBTB16*), using machine learning algorithms and survival analysis. The results were confirmed via qPCR of BC samples, and validated by another set of ML methods to measure the model accuracy and precision. To the best of our knowledge, the identified diagnostic and prognostic gene signatures are novel and have clinical potential.

Author Contributions: Conceptualization, Z.M., N.A. (Nesar Ahmad), N.A. (Nofe Alganmi), H.B. and S.K.; Data curation, Z.M., M.S.A., M.S.I., N.A. (Nesar Ahmad), N.A. (Nofe Alganmi), H.B. and S.K.; Formal analysis, M.S.A., M.S.I. and S.K.; Funding acquisition, Z.M., N.A. (Nofe Alganmi), H.B. and S.K.; Investigation, N.A. (Nesar Ahmad), N.A. (Nofe Alganmi), H.B. and S.K.; Methodology, M.S.A., M.S.I. and S.K.; Project administration, Z.M., N.A. (Nofe Alganmi), H.B. and S.K.; Resources, Z.M., N.A. (Nofe Alganmi), M.H.A.-Q. and S.K.; Software, M.S.A., M.S.I. and S.K.; Supervision, N.A. (Nesar Ahmad), M.H.A.-Q. and S.K.; Validation, Z.M. and S.K.; Visualization, N.A. (Nesar Ahmad) and M.H.A.-Q.; Writing—original draft, Z.M., M.S.A., M.S.I. and S.K.; Writing—review and editing, N.A. (Nesar Ahmad), N.A. (Nofe Alganmi), H.B. and M.H.A.-Q. All authors have read and agreed to the published version of the manuscript.

Funding: Deputyship for Research and Innovation, Ministry of Education and King Abdulaziz University, Saudi Arabia (project number IFPRC-125-141-2020).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the CEGMR bioethical committee (approval number 16-CEGMR-bioeth-2022, dated 13 October 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All data are publicly available, and their web links are mentioned in the Methods Section. However, additional information can be provided by the authors upon reasonable request.

Acknowledgments: The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number IFPRC-125-141-2020, and King Abdulaziz University, DSR, Jeddah, Saudi Arabia. The authors also thank the AZIZ-HPC facility of King Abdulaziz University, Jeddah, Saudi Arabia for computational support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Karim, S.; Al-Kharraz, M.; Mirza, Z.; Noureldin, H.; Abusamara, H.; Alganmi, N.; Merdad, A.; Jastaniah, S.; Kumar, S.; Rasool, M.; et al. Development of “Biosearch System” for biobank management and storage of disease associated genetic information. *J. King Saud Univ.—Sci.* **2022**, *34*, 101760. [[CrossRef](#)]
2. Ramaswamy, S.; Tamayo, P.; Rifkin, R.; Mukherjee, S.; Yeang, C.H.; Angelo, M.; Ladd, C.; Reich, M.; Latulippe, E.; Mesirov, J.P.; et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 15149–15154. [[CrossRef](#)] [[PubMed](#)]
3. Arnold, M.; Morgan, E.; Rungay, H.; Mafra, A.; Singh, D.; Laversanne, M.; Vignat, J.; Gralow, J.R.; Cardoso, F.; Siesling, S.; et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast* **2022**, *66*, 15–23. [[CrossRef](#)] [[PubMed](#)]
4. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [[CrossRef](#)] [[PubMed](#)]
5. Khan, A.; Rehman, Z.; Hashmi, H.F.; Khan, A.A.; Junaid, M.; Sayaf, A.M.; Ali, S.S.; Hassan, F.U.; Heng, W.; Wei, D.Q. An Integrated Systems Biology and Network-Based Approaches to Identify Novel Biomarkers in Breast Cancer Cell Lines Using Gene Expression Data. *Interdiscip. Sci.* **2020**, *12*, 155–168. [[CrossRef](#)] [[PubMed](#)]
6. Abd-Elnaby, M.; Alfonse, M.; Roushdy, M. Classification of breast cancer using microarray gene expression data: A survey. *J. Biomed. Inform.* **2021**, *117*, 103764. [[CrossRef](#)]
7. Makary, M.A.; Daniel, M. Medical error—The third leading cause of death in the US. *BMJ* **2016**, *353*, i2139. [[CrossRef](#)]

8. Karim, S.; Iqbal, M.S.; Ahmad, N.; Ansari, M.S.; Mirza, Z.; Merdad, A.; Jastaniah, S.D.; Kumar, S. Gene expression study of breast cancer using Welch Satterthwaite t-test, Kaplan-Meier estimator plot and Huber loss robust regression model. *J. King Saud Univ.—Sci.* **2023**, *35*, 102447. [[CrossRef](#)]
9. Schena, M.; Shalon, D.; Davis, R.W.; Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**, *270*, 467–470. [[CrossRef](#)]
10. Qing, T.; Karn, T.; Rozenblit, M.; Foldi, J.; Marczyk, M.; Shan, N.L.; Blenman, K.; Holtrich, U.; Kalinsky, K.; Meric-Bernstam, F.; et al. Molecular differences between younger versus older ER-positive and HER2-negative breast cancers. *NPJ Breast Cancer* **2022**, *8*, 119. [[CrossRef](#)]
11. Karim, S.; Merdad, A.; Schulten, H.J.; Jayapal, M.; Dallol, A.; Buhmeida, A.; Al-Thubaity, F.; Mirza, Z.; Gari, M.A.; Chaudhary, A.G.; et al. Low expression of leptin and its association with breast cancer: A transcriptomic study. *Oncol. Rep.* **2016**, *36*, 43–48. [[CrossRef](#)] [[PubMed](#)]
12. Merdad, A.; Karim, S.; Schulten, H.J.; Dallol, A.; Buhmeida, A.; Al-Thubaity, F.; Gari, M.A.; Chaudhary, A.G.; Abuzenadah, A.M.; Al-Qahtani, M.H. Expression of matrix metalloproteinases (MMPs) in primary human breast cancer: MMP-9 as a potential biomarker for cancer invasion and metastasis. *Anticancer Res.* **2014**, *34*, 1355–1366. [[PubMed](#)]
13. van de Vijver, M.J.; He, Y.D.; van't Veer, L.J.; Dai, H.; Hart, A.A.; Voskuil, D.W.; Schreiber, G.J.; Peterse, J.L.; Roberts, C.; Marton, M.J.; et al. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **2002**, *347*, 1999–2009. [[CrossRef](#)] [[PubMed](#)]
14. Schulten, H.J.; Bangash, M.; Karim, S.; Dallol, A.; Hussein, D.; Merdad, A.; Al-Thoubaity, F.K.; Al-Maghrabi, J.; Jamal, A.; Al-Ghamdi, F.; et al. Comprehensive molecular biomarker identification in breast cancer brain metastases. *J. Transl. Med.* **2017**, *15*, 269. [[CrossRef](#)] [[PubMed](#)]
15. Perou, C.M.; Sørlie, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; Rees, C.A.; Pollack, J.R.; Ross, D.T.; Johnsen, H.; Akslen, L.A.; et al. Molecular portraits of human breast tumours. *Nature* **2000**, *406*, 747–752. [[CrossRef](#)] [[PubMed](#)]
16. Khan, J.; Wei, J.S.; Ringnér, M.; Saal, L.H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C.R.; Peterson, C.; et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **2001**, *7*, 673–679. [[CrossRef](#)]
17. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **2012**, *490*, 61–70. [[CrossRef](#)]
18. Slodkowska, E.A.; Ross, J.S. MammaPrint 70-gene signature: Another milestone in personalized medical care for breast cancer patients. *Expert Rev. Mol. Diagn.* **2009**, *9*, 417–422. [[CrossRef](#)]
19. van 't Veer, L.J.; Dai, H.; van de Vijver, M.J.; He, Y.D.; Hart, A.A.; Mao, M.; Peterse, H.L.; van der Kooy, K.; Marton, M.J.; Witteveen, A.T.; et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **2002**, *415*, 530–536. [[CrossRef](#)]
20. Qian, Y.; Daza, J.; Itzel, T.; Betge, J.; Zhan, T.; Marmé, F.; Teufel, A. Prognostic Cancer Gene Expression Signatures: Current Status and Challenges. *Cells* **2021**, *10*, 648. [[CrossRef](#)]
21. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531–537. [[CrossRef](#)]
22. Sotiriou, C.; Pusztai, L. Gene-expression signatures in breast cancer. *N. Engl. J. Med.* **2009**, *360*, 790–800. [[CrossRef](#)]
23. Smyth, G.K. limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*; Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A., Dudoit, S., Eds.; Springer: New York, NY, USA, 2005; pp. 397–420.
24. Wang, G.; Muschelli, J.; Lindquist, M.A. Moderated *t*-tests for group-level fMRI analysis. *NeuroImage* **2021**, *237*, 118141. [[CrossRef](#)]
25. Sherman, B.T.; Hao, M.; Qiu, J.; Jiao, X.; Baseler, M.W.; Lane, H.C.; Imamichi, T.; Chang, W. DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **2022**, *50*, W216–W221. [[CrossRef](#)]
26. Liao, Y.; Wang, J.; Jaehnig, E.J.; Shi, Z.; Zhang, B. WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **2019**, *47*, W199–W205. [[CrossRef](#)]
27. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [[CrossRef](#)]
28. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
29. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [[CrossRef](#)]
30. Goldberg, D.E.; Holland, J.H. Genetic Algorithms and Machine Learning. *Mach. Learn.* **1988**, *3*, 95–99. [[CrossRef](#)]
31. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
32. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
33. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [[CrossRef](#)] [[PubMed](#)]
34. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [[CrossRef](#)]
35. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [[CrossRef](#)]
36. Iqbal, M.S.; Ahmad, N.; Mirza, Z.; Karim, S. Gene expression and survival analysis study of KIAA0101 gene revealed its prognostic and diagnostic importance in breast cancer. *Vegetos* **2023**, *36*, 249–258. [[CrossRef](#)]

37. Lániczky, A.; Gyórfy, B. Web-Based Survival Analysis Tool Tailored for Medical Research (KMplot): Development and Implementation. *J. Med. Internet Res.* **2021**, *23*, e27633. [[CrossRef](#)] [[PubMed](#)]
38. Nicolini, A.; Ferrari, P.; Duffy, M.J. Prognostic and predictive biomarkers in breast cancer: Past, present and future. *Semin. Cancer Biol.* **2018**, *52*, 56–73. [[CrossRef](#)]
39. Nair, M.; Sandhu, S.S.; Sharma, A.K. Cancer molecular markers: A guide to cancer detection and management. *Semin. Cancer Biol.* **2018**, *52*, 39–55. [[CrossRef](#)]
40. Senkus, E.; Kyriakides, S.; Ohno, S.; Penault-Llorca, F.; Poortmans, P.; Rutgers, E.; Zackrisson, S.; Cardoso, F. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **2015**, *26* (Suppl. 5), v8–v30. [[CrossRef](#)]
41. Chibon, F. Cancer gene expression signatures—The rise and fall? *Eur. J. Cancer* **2013**, *49*, 2000–2009. [[CrossRef](#)]
42. Kanathezath, A.; Chembra, V.; Padingare Variyath, K.S.; Nair, G.G. Identification of Biomarkers and Functional Modules from Genomic Data in Stage-wise Breast Cancer. *Curr. Bioinform.* **2021**, *16*, 722–733. [[CrossRef](#)]
43. Zhang, S.; Jiang, H.; Gao, B.; Yang, W.; Wang, G. Identification of Diagnostic Markers for Breast Cancer Based on Differential Gene Expression and Pathway Network. *Front. Cell Dev. Biol.* **2021**, *9*, 811585. [[CrossRef](#)]
44. Bao, S.; He, G. Identification of Key Genes and Key Pathways in Breast Cancer Based on Machine Learning. *Med. Sci. Monit.* **2022**, *28*, e935515. [[CrossRef](#)]
45. Dehdar, S.; Salimifard, K.; Mohammadi, R.; Marzban, M.; Saadatmand, S.; Fararouei, M.; Dianati-Nasab, M. Applications of different machine learning approaches in prediction of breast cancer diagnosis delay. *Front. Oncol.* **2023**, *13*, 1103369. [[CrossRef](#)]
46. Deng, J.-L.; Xu, Y.-h.; Wang, G. Identification of Potential Crucial Genes and Key Pathways in Breast Cancer Using Bioinformatic Analysis. *Front. Genet.* **2019**, *10*, 695. [[CrossRef](#)]
47. Joglekar, A.P.; Kukreja, A.A. How Kinetochore Architecture Shapes the Mechanisms of Its Function. *Curr. Biol.* **2017**, *27*, R816–R824. [[CrossRef](#)]
48. Cairo, G.; Lacefield, S. Establishing correct kinetochore-microtubule attachments in mitosis and meiosis. *Essays Biochem.* **2020**, *64*, 277–287.
49. Su, T.; Qin, X.-Y.; Dohmae, N.; Wei, F.; Furutani, Y.; Kojima, S.; Yu, W. Inhibition of Ganglioside Synthesis Suppressed Liver Cancer Cell Proliferation through Targeting Kinetochore Metaphase Signaling. *Metabolites* **2021**, *11*, 167. [[CrossRef](#)]
50. Carnero, A.; Blanco-Aparicio, C.; Renner, O.; Link, W.; Leal, J.F. The PTEN/PI3K/AKT signalling pathway in cancer, therapeutic implications. *Curr. Cancer Drug Targets* **2008**, *8*, 187–198. [[CrossRef](#)]
51. Carnero, A.; Paramio, J.M. The PTEN/PI3K/AKT Pathway in vivo, Cancer Mouse Models. *Front. Oncol.* **2014**, *4*, 252. [[CrossRef](#)]
52. Georgescu, M.M. PTEN Tumor Suppressor Network in PI3K-Akt Pathway Control. *Genes Cancer* **2010**, *1*, 1170–1177. [[CrossRef](#)]
53. Zhang, H. Molecular signaling and genetic pathways of senescence: Its role in tumorigenesis and aging. *J. Cell. Physiol.* **2007**, *210*, 567–574. [[CrossRef](#)] [[PubMed](#)]
54. Rayess, H.; Wang, M.B.; Srivatsan, E.S. Cellular senescence and tumor suppressor gene p16. *Int. J. Cancer* **2012**, *130*, 1715–1725. [[CrossRef](#)]
55. Bernardes de Jesus, B.; Blasco, M.A. Telomerase at the intersection of cancer and aging. *Trends Genet.* **2013**, *29*, 513–520. [[CrossRef](#)] [[PubMed](#)]
56. Ou, H.-L.; Hoffmann, R.; González-López, C.; Doherty, G.J.; Korkola, J.E.; Muñoz-Espín, D. Cellular senescence in cancer: From mechanisms to detection. *Mol. Oncol.* **2021**, *15*, 2634–2671. [[CrossRef](#)]
57. Sarkisian, C.J.; Keister, B.A.; Stairs, D.B.; Boxer, R.B.; Moody, S.E.; Chodosh, L.A. Dose-dependent oncogene-induced senescence in vivo and its evasion during mammary tumorigenesis. *Nat. Cell Biol.* **2007**, *9*, 493–505. [[CrossRef](#)] [[PubMed](#)]
58. Arena, P.; Basile, A.; Bucolo, M.; Fortuna, L. Image processing for medical diagnosis using CNN. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2003**, *497*, 174–178. [[CrossRef](#)]
59. Speiser, J.L.; Miller, M.E.; Tooze, J.; Ip, E. A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. *Expert Syst. Appl.* **2019**, *134*, 93–101. [[CrossRef](#)]
60. Chen, Y.; Zheng, W.; Li, W.; Huang, Y. Large group activity security risk assessment and risk early warning based on random forest algorithm. *Pattern Recognit. Lett.* **2021**, *144*, 1–5. [[CrossRef](#)]
61. Lee, S.; Kwon, S.; Kim, Y. A modified local quadratic approximation algorithm for penalized optimization problems. *Comput. Stat. Data Anal.* **2016**, *94*, 275–286. [[CrossRef](#)]
62. Koul, N.; Manvi, S.S. A Scheme for Feature Selection from Gene Expression Data using Recursive Feature Elimination with Cross Validation and Unsupervised Deep Belief Network Classifier. In Proceedings of the 2019 3rd International Conference on Computing and Communications Technologies (ICCT), Chennai, India, 21–22 February 2019; pp. 31–36.
63. Brownlee, J. *Deep Learning with Time Series Forecasting: Machine Learning Mastery*; San Juan, PR, USA, 2020; Volume 2023.
64. Ranstam, J.; Cook, J.A. LASSO regression. *Br. J. Surg.* **2018**, *105*, 1348. [[CrossRef](#)]
65. McEligot, A.J.; Poynor, V.; Sharma, R.; Panangadan, A. Logistic LASSO Regression for Dietary Intakes and Breast Cancer. *Nutrients* **2020**, *12*, 2652. [[CrossRef](#)]
66. Hastie, T.; Tibshirani, R.; Friedman, J. Support Vector Machines and Flexible Discriminants. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009; pp. 417–458.
67. Katoch, S.; Chauhan, S.S.; Kumar, V. A review on genetic algorithm: Past, present, and future. *Multimed. Tools Appl.* **2021**, *80*, 8091–8126. [[CrossRef](#)]

68. Puppe, J.; Seifert, T.; Eichler, C.; Pilch, H.; Mallmann, P.; Malter, W. Genomic Signatures in Luminal Breast Cancer. *Breast Care* **2020**, *15*, 355–365. [[CrossRef](#)]
69. Varnier, R.; Sajous, C.; de Talhouet, S.; Smentek, C.; Péron, J.; You, B.; Reverdy, T.; Freyer, G. Using Breast Cancer Gene Expression Signatures in Clinical Practice: Unsolved Issues, Ongoing Trials and Future Perspectives. *Cancers* **2021**, *13*, 4840. [[CrossRef](#)]
70. Nasser, M.; Yusof, U.K. Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. *Diagnostics* **2023**, *13*, 161. [[CrossRef](#)]
71. Thalor, A.; Kumar Joon, H.; Singh, G.; Roy, S.; Gupta, D. Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 1618–1631. [[CrossRef](#)]
72. Taghizadeh, E.; Heydarheydari, S.; Saberi, A.; JafarpourNesheli, S.; Rezaei, S.M. Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC Bioinform.* **2022**, *23*, 410. [[CrossRef](#)]
73. Li, Q.; Yang, H.; Wang, P.; Liu, X.; Lv, K.; Ye, M. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. *J. Transl. Med.* **2022**, *20*, 177. [[CrossRef](#)]
74. Kurian, B.; Jyothi, V.L. Comparative Analysis of Machine Learning Methods for Breast Cancer Classification in Genetic Sequences. *J. Environ. Public Health* **2022**, *2022*, 7199290. [[CrossRef](#)]
75. Tabl, A.A.; Alkhateeb, A.; ElMaraghy, W.; Rueda, L.; Ngom, A. A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. *Front. Genet.* **2019**, *10*, 256. [[CrossRef](#)]
76. Kim, B.-C.; Kim, J.; Lim, I.; Kim, D.H.; Lim, S.M.; Woo, S.-K. Machine Learning Model for Lymph Node Metastasis Prediction in Breast Cancer Using Random Forest Algorithm and Mitochondrial Metabolism Hub Genes. *Appl. Sci.* **2021**, *11*, 2897. [[CrossRef](#)]
77. Sieuwerts, A.M.; Look, M.P.; Meijer-van Gelder, M.E.; Timmermans, M.; Trapman, A.M.A.C.; Garcia, R.R.; Arnold, M.; Goedheer, A.J.W.; de Weerd, V.; Portengen, H.; et al. Which Cyclin E Prevails as Prognostic Marker for Breast Cancer? Results from a Retrospective Study Involving 635 Lymph Node–Negative Breast Cancer Patients. *Clin. Cancer Res.* **2006**, *12*, 3319–3328. [[CrossRef](#)] [[PubMed](#)]
78. Liu, N.-Q.; Cao, W.-H.; Wang, X.; Chen, J.; Nie, J. Cyclin genes as potential novel prognostic biomarkers and therapeutic targets in breast cancer. *Oncol. Lett.* **2022**, *24*, 374. [[CrossRef](#)] [[PubMed](#)]
79. Liu, R.; Guo, C.-X.; Zhou, H.-H. Network-based approach to identify prognostic biomarkers for estrogen receptor–positive breast cancer treatment with tamoxifen. *Cancer Biol. Ther.* **2015**, *16*, 317–324. [[CrossRef](#)]
80. Weng, Y.; Liang, W.; Ji, Y.; Li, Z.; Jia, R.; Liang, Y.; Ning, P.; Xu, Y. Key Genes and Prognostic Analysis in HER2+ Breast Cancer. *Technol. Cancer Res. Treat.* **2021**, *20*, 1533033820983298. [[CrossRef](#)]
81. Jiang, Y.; Liu, Y.; Tan, X.; Yu, S.; Luo, J. TPX2 as a Novel Prognostic Indicator and Promising Therapeutic Target in Triple-negative Breast Cancer. *Clin. Breast Cancer* **2019**, *19*, 450–455. [[CrossRef](#)]
82. Abuderman, A.; Harb, O.; Gertallah, L. Prognostic and clinicopathological values of tissue expression of MFAP5 and ITM2A in triple-negative breast cancer: An immunohistochemical study. *Contemp. Oncol./Współczesna Onkol.* **2020**, *24*, 87–95. [[CrossRef](#)]
83. Chen, D.; Sun, Y.; Wei, Y.; Zhang, P.; Rezaeian, A.H.; Teruya-Feldstein, J.; Gupta, S.; Liang, H.; Lin, H.-K.; Hung, M.-C.; et al. LIFR is a breast cancer metastasis suppressor upstream of the Hippo-YAP pathway and a prognostic marker. *Nat. Med.* **2012**, *18*, 1511–1517. [[CrossRef](#)]
84. van Ijzendoorn, D.G.P.; Szuhai, K.; Briare-de Bruijn, I.H.; Kostine, M.; Kuijjer, M.L.; Bovée, J.V.M.G. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS Comput. Biol.* **2019**, *15*, e1006826. [[CrossRef](#)]
85. He, J.; Wu, M.; Xiong, L.; Gong, Y.; Yu, R.; Peng, W.; Li, L.; Li, L.; Tian, S.; Wang, Y.; et al. BTB/POZ zinc finger protein ZBTB16 inhibits breast cancer proliferation and metastasis through upregulating ZBTB28 and antagonizing BCL6/ZBTB27. *Clin. Epigenet.* **2020**, *12*, 82. [[CrossRef](#)]
86. Hao, M.; Liu, W.; Ding, C.; Peng, X.; Zhang, Y.; Chen, H.; Dong, L.; Liu, X.; Zhao, Y.; Chen, X.; et al. Identification of hub genes and small molecule therapeutic drugs related to breast cancer with comprehensive bioinformatics analysis. *PeerJ* **2020**, *8*, e9946. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.