Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/40/20241235

Insights on the clinical, genetic, and mutation risk factors underlying the prognosis of breast cancer using machine learning models

Yundi Chen

College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei, China

cindyt@capstone-research.com

Abstract. Breast cancer is one of the most prevalent tumors in the world, making it essential to identify important features for the prognosis. Various methods, including statistical tests and machine learning, have been employed to capture relative attributes for survival prediction. However, most prior studies have focused solely on clinical factors, without considering genetic factors. In this study, a dataset comprising clinical features, gene expression, mutation attributes, and survival status of 1882 patients was involved to assess important features for breast cancer prognosis. Statistic tests were applied first to identify distribution differences and correlation significance among different features. Afterward, predictive models were trained. The results indicated that "age at diagnosis", "lymph nodes examined positive", and "Nottingham prognostic index" were the top three important features. Genes including *HSD17B11*, *JAK1*, and *STAT5A*, as well as mutations, including mutations in *GATA3*, *TP53*, and *MUC16*, also emerged as relative features. Additionally, Gradient Boosting Decision Trees outperformed three other models with an AUC-ROC of 0.75. These findings shed light on the further identification of not only important clinical attributes but also molecular markers for breast cancer prognosis.

Keywords: Breast cancer, survival prediction, Molecular markers, Machine learning, Classification.

1. Introduction

Breast cancer is one of the most common malignant tumors worldwide, accounting for 36% of oncological cases [1]. It was estimated that in 2023, approximately 31% of newly diagnosed cancer cases for women in the United States would be breast cancer [2]. Despite increasingly effective diagnosis and therapies, breast cancer remains the second leading cause of cancer-related death in women worldwide [3], with it being the leading cause among women aged 20 to 49 years [2].

The traditional classification of breast cancer includes "histological categorization", "tumor stage", "neoplasm histologic grade", and "receptor status". In terms of histological categorization, breast cancer can be classified as in-situ carcinoma or invasive carcinoma, depending on whether the tumor is limited to the breast epithelial layer or has invaded surrounding areas. For invasive carcinoma, invasive ductal carcinoma (IDC) accounts for 50% to 80% of newly diagnosed cases, and the rest are invasive lobular carcinomas (ILC) [4]. "Tumor stage" is determined by the TNM Classification and categorizes patients

DOI: 10.54254/2753-8818/40/20241235

into stages 0 to 4. The staging system involves three mandatory parameters: T, represents the size of the primary tumor; N, describes the degree of regional spread; M, indicates the presence of distant metastasis. Stage is the most important factor to consider when choosing treatments. Additionally, depending on the morphological assessment of tumor tissue under the microscope, "neoplasm histologic grade" assigns tumors an integer score ranging from 1 to 3 [5]. As for "receptor status", it is determined by assessment of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), categorizing patients into different positive or negative types based on the expression of these prognostic markers.

The molecular classification includes "intrinsic subtypes", "integrative clusters", and "gene mutations". "Intrinsic subtypes" can be identified through gene expression profiling and immunohistochemistry, by determining expression patterns of four biomarkers, including ER, PR, HER2, and cell proliferation regulator (Ki-67), which categorizes patients into five subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like, and Claudin-low) [6]. Among all categories, Luminal A is the most common case, constituting around half of the newly diagnosed cases [7], and has the best prognosis due to its relatively low grade [8]. Although HER2-enriched tumor is often high-grade, its response to anti-HER2-targeted therapy is aggressive, which significantly improves the outcome [8]. "Integrative clusters" assign tumors into IntClust 1 to 10, based on aberration of gene expressions. Among all clusters, IntClust 5 has the poorest outcome [9]. Mutations are infrequent in breast cancers. Among all the mutation types, *PIK3CA* and *TP53* dominate, with incidences of 40.1% and 35.4% respectively [10].

Many risk factors could contribute to breast cancer development and progression. Family history, for example, the presence of *BRCA1/2* mutation, contributes to gene susceptibility [11]. Unhealthy lifestyle such as a high-fat diet, lack of physical activity, and smoking could lead to an increased risk of breast cancer, by disturbing hormone levels and inducing gene mutations [12]. Late marriage and late childbirth also play a negative role by exerting influence on breast tissue differentiation and genotoxicity by estrogen [12,13].

Previously, many studies had employed machine learning methods for survival prediction of breast cancer. The relationship between clinical attributes and pathological characteristics in the survival of breast cancer patients has been analyzed [14]. Additionally, "tumor size", "stage", "age at diagnosis", "total axillary lymph node removed", and "number of positive lymph nodes" had been identified as the most relative factors involved in breast cancer prognosis [15]. Moreover, a recent study established a 5-year survival predictive model using machine learning with clinical features of "cancer stage", "tumor size", "diagnosis age", "surgery", and "body mass index" [16]. Furthermore, Random Forest was recommended for breast cancer survival prediction [17]. In addition, some studies involved deep learning methods to further enhance the prediction performance with clinical records [15,18]. However, these previous studies primarily focused on clinical features and paid limited attention to genetic features and mutation.

Therefore, in the present study, a dataset with clinical attributes, genetic features, and mutations was involved to gain a more comprehensive view of important features and molecular markers for breast cancer prognosis. Multiple statistical tests were employed for the dataset, followed by the training of four machine learning models. The performance and predictions of the models were analyzed and compared. The results indicated that "age at diagnosis", "lymph nodes examined positive", and "Nottingham prognostic index" were among the top important features. Genes including *HSD17B11*, *JAK1*, and *STAT5A*, as well as mutations, including mutations in *GATA3*, *TP53*, and *MUC16*, also emerged as important predictive features of cancer survival.

2. Methods

2.1. Overview of dataset

The dataset had 687 columns and 1882 rows. Each column represented a certain feature while every row corresponded to an individual patient. The dataset included three kinds of attributes: clinical indicators

DOI: 10.54254/2753-8818/40/20241235

with 25 columns, gene expressions with 489 columns, and gene mutations with 173 columns. The data of gene expressions and mutation was complete with no missing values. However, 11 of the clinical features were incomplete. Specifically, the "tumor stage" had 26% missing values, the "3-gene classifier subtype" had 11%, and the remaining features had missing values ranging from 0.8% to 6%. In data visualization and statistical tests, all missing values were removed. As for model training, missing values were filled with the mean or mode of the test set, depending on whether the feature was numeric or categorical.

2.2. Features description

2.2.1. Numeric clinical attributes. There were 6 numeric clinical attributes in total: "age at diagnosis" (yr), "lymph nodes examined positive", "mutation count", "overall survival months", "tumor size" (mm), and "Nottingham prognostic index". "Age at diagnosis" refers to the age at which the patient was diagnosed with cancer. "Lymph nodes examined positive" refers to the number of lymph nodes taken during the surgery that were determined to contain cancer cells. "Mutation count" refers to the number of genes with relevant mutations. "Overall survival months" was defined as the time from randomization to death. "Tumor size" was obtained through imaging techniques. "Nottingham prognostic index" was a prognostic measure calculated based on the size of the tumor, the number of lymph nodes involved, and the grade of the tumor [19].

2.2.2. Categorical clinical attributes. Categorical clinical data included 19 attributes in total: "cancer type detailed", "type of breast surgery", "primary tumor laterality", "inferred menopausal state", "tumor stage", "chemotherapy", "hormone therapy", "radio therapy", "cellularity", "neoplasm histologic grade", "tumor other histologic subtype", "pam50 + claudin-low subtype", "PR status", "ER status", "ER status measured by IHC", "HER2 status", "HER2 status measured by SNP6", "3-gene classifier subtype", "integrative cluster".

"Cancer type detailed" included four detailed breast cancer types: breast invasive ductal carcinoma (IDC), breast mixed ductal and lobular carcinoma (MDLC), breast invasive lobular carcinoma (ILC), and breast invasive mixed mucinous carcinoma (IMMC). "Primary tumor laterality" indicates whether the primary tumor originates from the right breast or the left breast. "Inferred menopausal state" employs "post" or "pre" to indicate whether the patient was postmenopausal or not. "Tumor stage" took a value from 1 to 4 based on the involvement of surrounding structures, lymph nodes, and the spread distance. "Type of breast surgery" contained two types of breast cancer surgery: mastectomy and breast conserving. While the mastectomy removed all breast tissues, the breast-conserving only removed the part that had cancer. "Chemotherapy", "Hormone therapy" and "Radio therapy" were binary classification categories that employed the values 0 and 1 to denote whether patients had or had not undergone the specific therapy respectively. "Cellularity" divided patients into three categories: high, moderate, and low, based on the amount and cluster arrangement of tumor cells in the specimen. "Neoplasm histologic grade" describes the extent to which tumor cells resemble normal cells in morphology. The value ranges from 1 to 3 in ascending order of aggressiveness. "Tumor other histologic subtype" classified patients into seven cancer types based on microscopic examination of cancer tissues: Ductal/NST, Mixed, Lobular, Tubular/cribriform, Mucinous, Medullary, and Other. "PAM50 + claudin-low subtype" was a combination of PAM50 and Claudin-low molecular subtypes. PAM50 was a tumor profiling test designed to prognosticate the likelihood of metastasis. It tested the expressions of a group of 50 genes from the tumor sample removed during the surgery and classified breast cancer into five molecular subtypes: Luminal A, Luminal B, human epidermal growth factor receptor 2 (HER2)-enriched, Basal-like, and Normal-like. Claudin-low molecular subtype was defined by low expression of claudins, enrichment for epithelial-to-mesenchymal transition (EMT), and tumor-initiating cell (TIC) features. "PR status" describes whether the cancer cells of the patient were positive or negative for progesterone receptor expression. "ER status" indicated whether subjects were positive or negative for estrogen receptor expression at the cellular level, while "ER status measured by IHC" corresponded to

DOI: 10.54254/2753-8818/40/20241235

the results obtained from immunohistochemistry. “HER2 status” referred to whether subjects were positive or negative for HER2 at the cellular level, while “HER2 status measured by SNP6” corresponded to the results obtained from SNP Array 6.0, a type of next-generation sequencing. “3-gene classifier subtype” classified patients in to “ER-/HER2-”, “ER+/HER2- High Prolif”, “ER+/HER2- Low Prolif” and “HER2+”, based on gene expressions. “Integrative cluster” refers to molecular subtypes of cancer based on some gene expression, including “1”, “2”, “3”, “4ER-”, “4ER+”, “5”, “6”, “7”, “8”, “9”, and “10”.

2.2.3. Genetic attributes. Genetic attributes in the dataset were z scores for mRNA expression of 331 genes, including BRCA1, BRCA2, PALB2, PTEN, TP53, ATM, CDH1, CHEK2, NBN, NF1, etc. The z score was calculated by dividing the difference between the expression in the tumor sample and the mean expression in the reference group (all samples in the study) by the standard deviation of expression in the reference group. The formula was as follows:

$$z = \frac{(expression\_in\_tumor\_sample - mean\_expression\_in\_reference\_sample)}{standard\_deviation\_of\_expression\_in\_reference\_sample}$$

2.2.4. Mutation attributes. Mutation attributes recorded whether a mutation occurs (indicated by 1) or not (indicated by 0) for 175 genes, including PIK3CA, TP53, MUC16, AHNAK2, KMT2C, SYNE1, GATA3, MAP3K1, AHNAK, DNAH11, etc.

2.2.5. Target attribute. “Overall survival” was used as the target attribute, in which 1 referred to the “survived” group, while 0 referred to the “died” group. The “survived” group contained 789 patients, taking up 42% of the total. The “died” group contained 1098 patients, taking up 58% of the total.

2.2.6. Statistical tests. For the numeric attributes, the Kolmogorov-Smirnov test was first conducted as a nonparametric method to determine if the two distributions across overall survival statuses significantly differ from each other. Then the Shapiro-Wilk test was carried out for normality validation. If the normality is validated, the Student’s t-test was applied to test the significance of the mean value difference between the two distributions. Otherwise, the Mann-Whitney U test was conducted for the significance of the median value difference. For the categorical attributes, the $\chi 2$ test was applied for distribution difference across overall survival statuses within each specific category. As most of the attributes in the study are not normally distributed, Spearman’s correlation was applied as the nonparametric approach for the correlation test.

2.3. Machine learning methods and performance matrices

2.3.1. Data pre-processing. First, the data set was randomly divided into a train set and a test set, in which the test set took a portion of 20%. Stratified sampling was employed to ensure the class distribution of the target variable, overall survival, was preserved in both the train and test sets.

2.3.2. Models involved. The modeling process included training in Logistic Regression (LR), Random Forest (RF), Gradient Boosting Decision Trees (GBDT), and Support Vector Classifier (SVC). Additionally, the five-fold cross-validation was incorporated in each classification.

2.3.3. Performance evaluation. After training, five matrices were used to determine model performance: accuracy, precision, recall, f1 score, and the Area Under the Receiver Operating Characteristics Curve (ROC-AUC). By comparing the prediction of the classifier with the target feature to which the sample truly belongs, outcomes were categorized into true positive (TP), true negative (TN), false positive (FP), and false negative (FN), comprising a confusion matrix. Accuracy is calculated as the ratio of correct predictions to the total number of predictions ($Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$). It provides an overall assessment of performance across all classes and is useful when classes are equally important and

DOI: 10.54254/2753-8818/40/20241235

balanced in quantity. Precision denotes the proportion of positive samples correctly predicted of a total number of sampled classifieds as positive ($Precision = \frac{TP}{TP + FP}$). It reflects how reliable the model is in classifying samples as positive. When the importance is strongly addressed on avoiding misclassifying negative cases as positive, the precision is a more suitable matrix to consider. Recall refers to the ratio of positive samples correctly predicted to the total number of true positive samples ($Recall = \frac{TP}{TP + FN}$). It reflects how many positive cases are correctly classified and disregards the prediction of negative cases. Recall is a more valuable matrix when the goal is to detect all positive samples. F1 score is the harmonic mean of the precision and the recall ($F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}$). It can be interpreted as the model's balanced capacity to both be accurate at the cases it predicts (precision) and be effective at capturing positive cases (recall). In the ROC, the true positive rate ($TPR = \frac{TP}{TP + FN}$) is plotted against the false positive rate ($FPR = \frac{TN}{TN + FP}$) under various thresholds. The higher the AUC is, the better the model is at distinguishing positive against negative labels.

2.3.4. Feature importance. Furthermore, the top 10 attributes with the highest significance in the model training were identified. In the case of RF and GBDT, the criterion was based on the value of feature importance, determined by impurity. For LG and SVC, it was based on the weight of each attribute in a linear model.

2.3.5. Principal Component Analysis. To reduce noise and dimensionality, Principal Component Analysis (PCA) was conducted in the end with standardized numeric clinical data and gene expressions. PCA is a type of unsupervised learning method that is similar to clustering, which reduces the number of variables by geometrically projecting them to principal components (PCs), regardless of samples' differences [20]. PCs are linear combinations of initial variables. Each PC is selected by the process of minimizing the distance between variables and their projection on the PC, while maximizing the variances of projected points, with the additional requirement that subsequent PCs have to be uncorrelated with previous ones [20]. In this approach, the first PC accounts for the largest explained variance. By evaluating the proportion of explained variance within each PC from PCA plots, insights into variable selections could be obtained and thus dimensionality is reduced [21].

3. Results

3.1. Statistics

3.1.1. Numeric Clinical Attributes. To start with, histogram and Kernel Density Estimation (KDE) are applied to visualize the distribution of numeric clinical features ("age at diagnosis", "lymph nodes examined positive", "mutation count", "tumor size", and "Nottingham prognostic index") across overall survival status, shown in figure 1. Additionally, the boxplot (figure 2) and violin plot (figure 3) are employed to show the distribution differences between the "survived" groups and the "died" groups.

To assess the distribution difference, the Kolmogorov-Smirnov test was conducted. The results revealed a significant difference between the two survival statuses within each of five numeric clinical features: "age at diagnosis" ("survived" = 60.41±13.01 yr, "died" = 63.87±13.13 yr, $p = 7.22 \times 10^{-23}$, Kolmogorov-Smirnov test), "lymph nodes examined positive" ("survived" = 1.18±2.73, "died" = 2.45±4.48, $p = 2.17 \times 10^{-4}$), "mutation count" ("survived" = 4.94±3.06, "died" = 5.91±4.28, $p = 5.07 \times 10^{-4}$), "tumor size" ("survived" = 22.95±11.39 mm, "died" = 28.40±16.95 mm, $p = 2.03 \times 10^{-9}$), and "Nottingham prognostic index" ("survived" = 3.93±0.99, "died" = 4.29±1.09, $p = 1.49 \times 10^{-9}$).

Following that, to study the normality of distributions, the Shapiro-Wilk test was conducted and the results showed that only the normality of the "survived" group within "age at diagnosis" was validated (statistic = 0.99, $p = 1.12 \times 10^{-2}$, Shapiro-Wilk test), consistent with figure 1. Consequently, the Mann-Whitney U test was applied to assess differences in median values. The results demonstrated that all five numeric clinical features have significantly different median values across overall survival statuses: "age

DOI: 10.54254/2753-8818/40/20241235

at diagnosis" (statistic = 9.24×10^{4}, p = 2.58×10^{-25}, Mann-Whitney U test), "lymph nodes examined positive" (statistic = 1.21×10^{5}, p = 1.29×10^{-7}), "mutation count" (statistic = 1.25×10^{5}, p = 5.67×10^{-5}), "tumor size" (statistic = 1.11×10^{5}, p = 1.51×10^{-11}), and "Nottingham prognostic index" (statistic = 1.13×10^{5}, p = 9.04×10^{-11}).

In addition, Spearman's correlation was applied to study the correlation among five numeric clinical features, which reveals that, except for "lymph nodes examined positive" and "Nottingham prognostic index" (coefficient = 0.77, p = 3.16×10^{-216}, Spearman's correlation), other features do not have significant correlation (coefficient > 0.6 and p < 0.05) with each other. Furthermore, a heatmap was employed to illustrate the correlation, shown in figure 4.

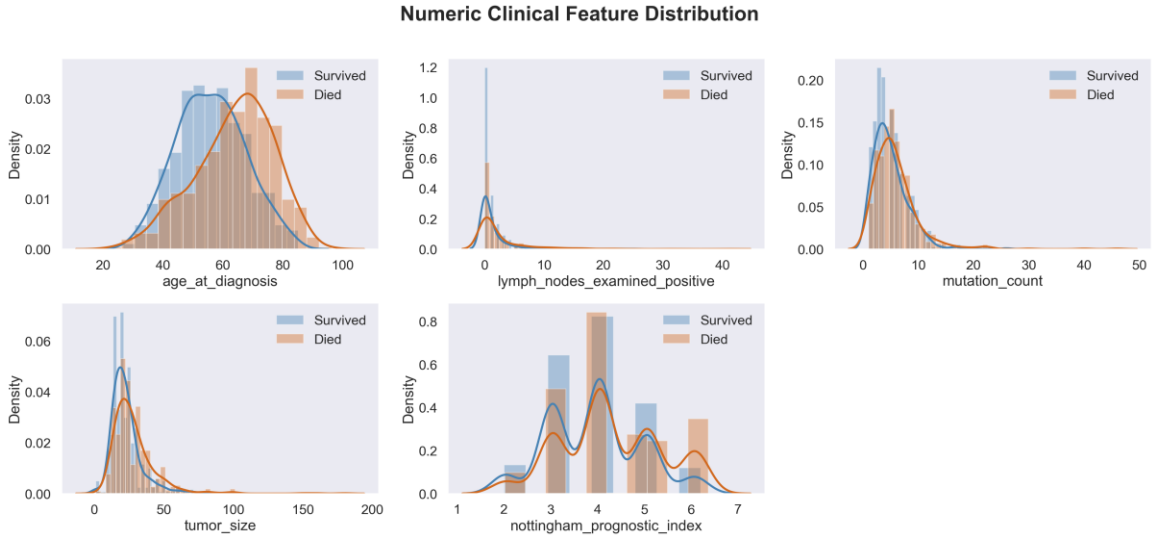

Figure 1. Distributions of numeric clinical features. The values on the vertical axes represent the frequency. The smooth KDE curves indicate the probability density distributions of variables. Blue refers to the "survived" group, while orange refers to the "died" group.

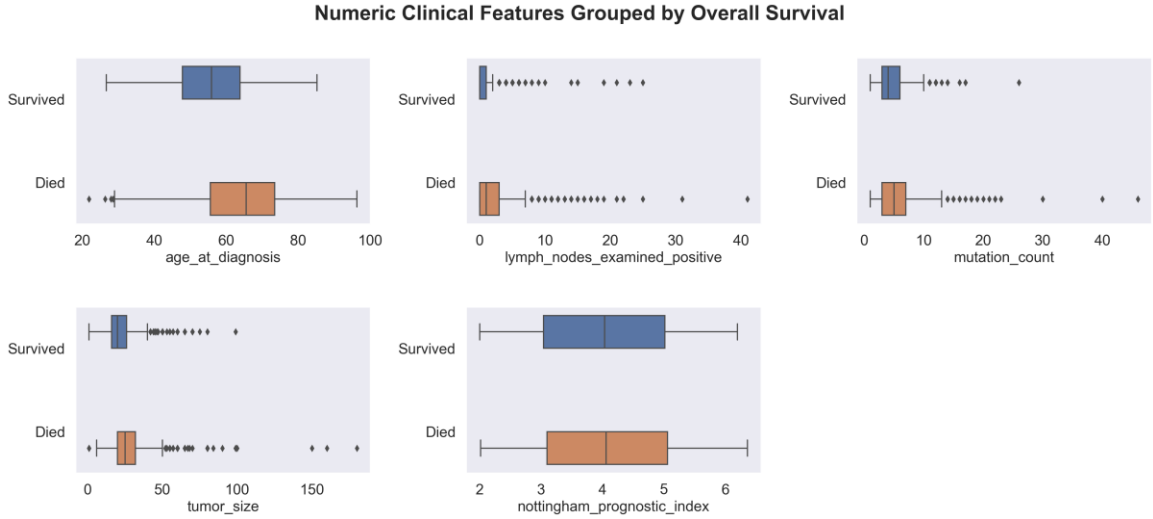

Figure 2. Boxplots of numeric clinical features. The horizontal axes present the values of different variables. Five vertical strings refer to whisker ($Q1 - 1.5$IQR), lower quartile (Q1, referring to the 25th percentile), median, upper quartile (Q3, referring to the 75th percentile), whisker ($Q3 + 1.5$IQR), from left to right. $IQR = Q3 - Q1$.

DOI: 10.54254/2753-8818/40/20241235

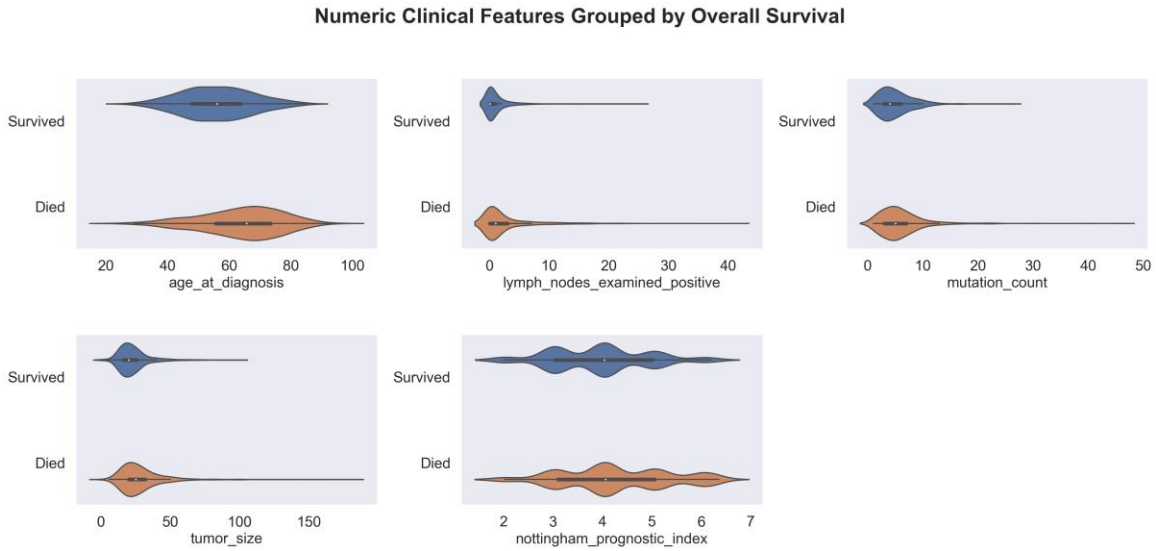

Figure 3. Violin plots of numeric clinical features. The horizontal axes present the values of different variables. Median values and quartiles of the "survived" group and the "died" group are displayed in each plot.

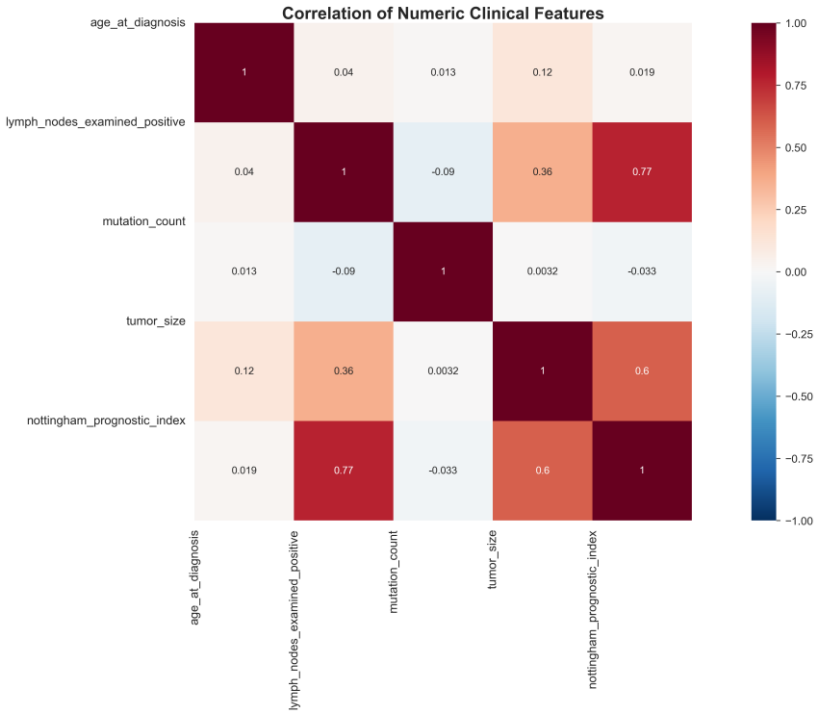

Figure 4. Heatmap of numeric clinical features. The horizontal and vertical axes present five variables. The number in each box refers to the correlation coefficient and the color indicates the strength of correlation.

3.1.2. Categorical Clinical Attributes. To illustrate the distribution of 19 categorical clinical attributes, a histogram was used to visualize the ratios of two survival statuses within each category of attributes (figure 5-7).

Following that, the $\chi 2$ test was conducted to assess the distribution difference across the overall survival statuses of each attribute. Results showed that 10 of 19 categorical clinical attributes had

DOI: 10.54254/2753-8818/40/20241235

significantly different distributions ($p < 0.05$) across overall survival statuses: “type of breast surgery”, “pam50 + claudin-low subtype”, “neoplasm histologic grade”, “tumor other histologic subtype”, “inferred menopausal state”, “integrative cluster”, “integrative cluster”, “primary tumor laterality”, “radio therapy”, “3-gene classifier subtype”, and “tumor stage”. Detailed data is listed in table 1.

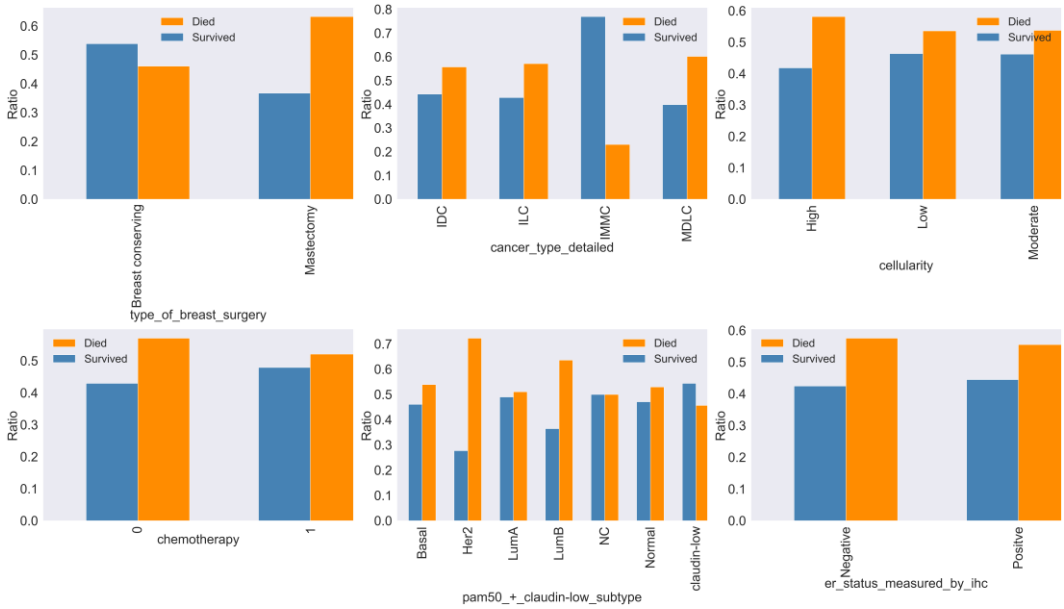

Figure 5. Ratio distribution of categorical clinical features (“type of breast surgery”, “cancer type detailed”, “cellularity”, “chemotherapy”, “pam50 + claudin-low subtype”, “ER status measured by IHC”). The horizontal axes list categories within each attribute and the vertical axes present ratios.

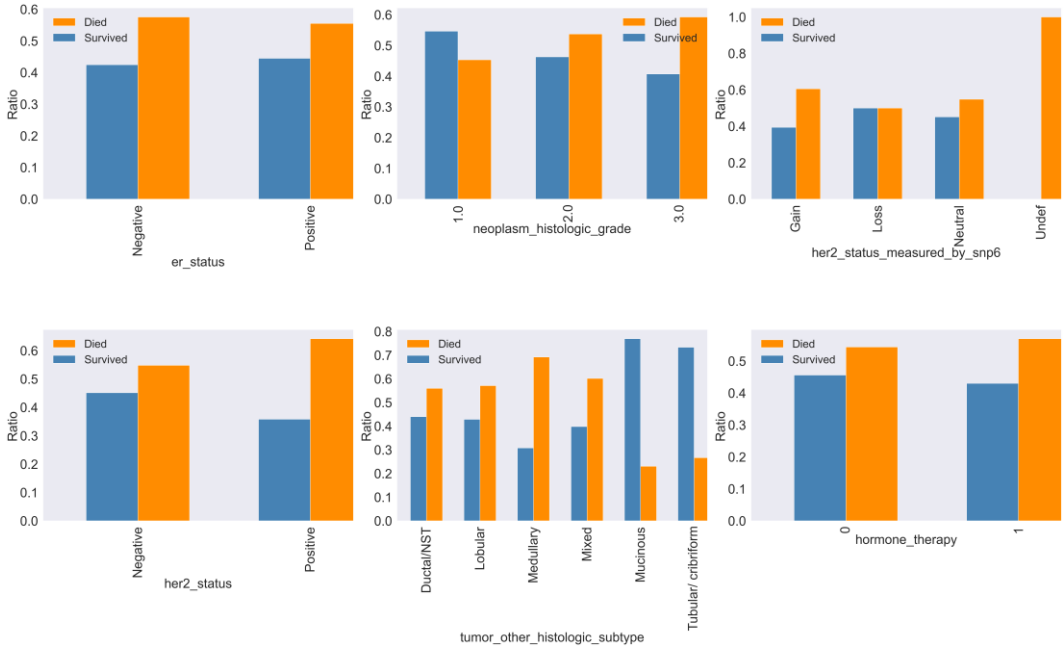

Figure 6. Ratio distribution of categorical clinical features (“ER status”, “neoplasm histologic grade”, “HER2 status measured by SNP6”, “HER2 status”, “tumor other histologic subtype”, “hormone therapy”).

DOI: 10.54254/2753-8818/40/20241235

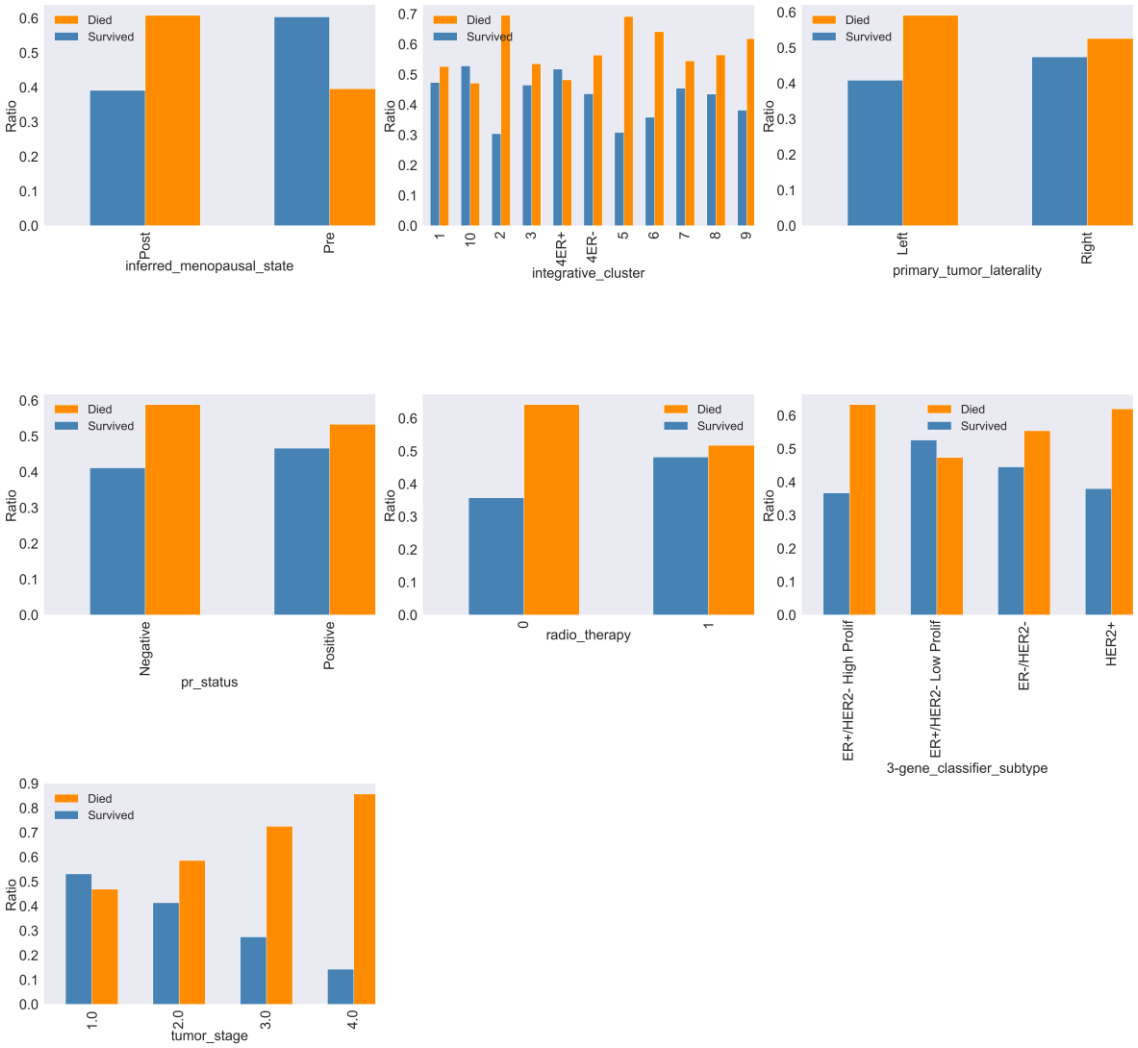

Figure 7. Ratio distribution of categorical clinical features ("inferred menopausal state", "integrative cluster", "primary tumor laterality", "PR status", "radio therapy", "3-gene classifier subtype", "tumor stage").

Table 1. $\chi 2$ and p values of categorical clinical attributes with significantly different distributions across survival status.

Attributes	$\chi 2$	P value
type of breast surgery	31.05	2.51×10^{-8}
pam50 + claudin-low subtype	26.54	1.77×10^{-4}
neoplasm histologic grade	7.34	2.55×10^{-2}
tumor other histologic subtype	12.88	2.45×10^{-2}
inferred menopausal state	34.46	4.36×10^{-9}
integrative cluster	21.68	1.68×10^{-2}
primary tumor laterality	4.44	3.50×10^{-2}
radio therapy	14.67	1.28×10^{-4}
3-gene classifier subtype	22.11	6.20×10^{-5}
tumor stage	26.76	6.60×10^{-6}

DOI: 10.54254/2753-8818/40/20241235

3.1.3. Genetic Attributes. To gain insight into distributions of genetic attributes, histogram and KDE were applied to the first nine genes (figure 8), which showed that some distributions seemed to be normal, while others were right-skewed or left-skewed.

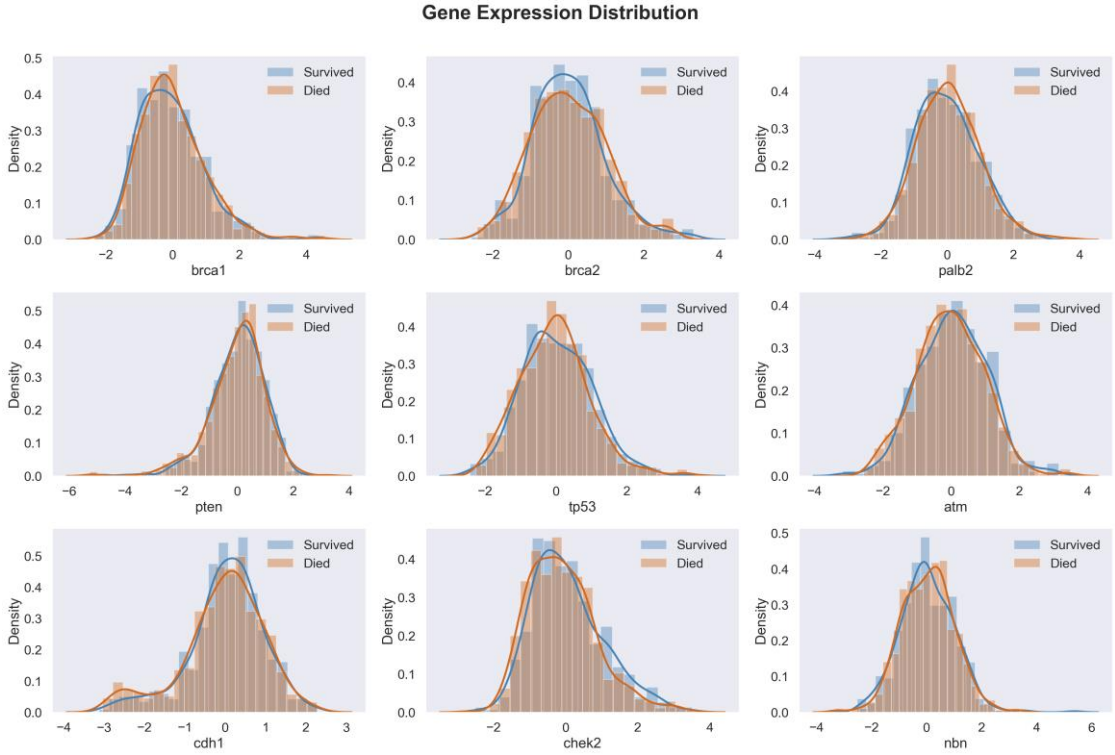

Figure 8. Distributions of first nine genetic attributes (*BRCA1, BRCA2, PALB2, PTEN, TP53, ATM, CDH1, CHEK2, NBN*).

Following that, the Shapiro-Wilk test was conducted. The results revealed that 53 genes in total were normally distributed ($p > 0.05$), including *HSD17B11* ("survived" = 0.31±0.89, "died" = -0.08±0.87, p of "survived" = 0.49, p of "died" = 0.28, Shapiro-Wilk test), *ATM* ("survived" = 0.05±0.99, "died" = -0.03±0.99, p of "survived" = 0.32, p of "died" = 0.08), *RAD50* ("survived" = -0.04±0.95, "died" = 0.04±1.05, p of "survived" = 0.87, p of "died" = 0.82,), *RB1* ("survived" = 0.18±0.92, "died" = 0.00±0.95, p of "survived" = 0.59, p of "died" = 0.21), *MYC* ("survived" = 0.25±0.92, "died" = -0.09±0.96, p of "survived" = 0.85, p of "died" = 0.85), etc..

For these 53 normally distributed genes, Student's t-test was applied to study the distribution difference across two survival statuses. Results revealed that 12 genes are significantly different in the mean value of expression across survival statuses ($p < 0.05$), including *HSD17B11* (statistic = -7.15, $p = 1.64\times10^{-12}$, Student's t-test), *RB1* (statistic = -3.21, $p = 1.36\times10^{-3}$), *MYC* (statistic = -5.92, $p = 4.45\times10^{-9}$), *STAT2* ("survived" = -0.12±0.92, "died" = 0.12±0.92, statistic = 4.21, $p = 2.76\times10^{-5}$), *TNK2* ("survived" = -0.14±0.80, "died" = 0.08±0.91, statistic = 4.19, $p = 3.05\times10^{-5}$), etc.

For other 436 non-normally distributed genes, the Mann-Whitney U test was applied to assess differences in median values. Results revealed that 184 genes showed significant difference ($p < 0.05$), including *STAT5A* ("survived" = 0.08±1.00, "died" = -0.23±0.90, statistic = 1.19×10^5, $p = 2.41\times10^{-7}$, Mann-Whitney U test), *JAK1* ("survived" = 0.26±0.90, "died" = -0.08±0.91, statistic = 1.13×10^5, $p = 1.15\times10^{-10}$, *NCOA3* ("survived" = -0.25±0.88, "died" = 0.08±0.98, statistic = 1.74×10^5, $p = 7.45\times10^{-8}$), *KIT* ("survived" = 0.18±0.96, "died" = -0.10±0.98, statistic = 1.18×10^5, $p = 9.66\times10^{-8}$), *CDK4* ("survived" = 0.11±1.00, "died" = 0.00±0.95, statistic = 1.33×10^5, $p = 1.45\times10^{-2}$), etc..

In total, 196 genes had significantly different distributions across overall survival statuses. For these significant genes, Spearman's correlation was carried out to study the correlation of genes. According

DOI: 10.54254/2753-8818/40/20241235

to the results, 42 pairs of genes were significantly correlated (coefficient > 0.6 and $p < 0.05$). The top five pairs of correlation were *AKR1C3* and *AKR1C4* (coefficient = 0.86 and $p = 3.04\times10^{-312}$, Spearman's correlation), *AKR1C2* and *AKR1C4* (coefficient = 0.84 and $p = 2.91\times 10^{-293}$), *E2F2* and *AURKA* (coefficient =0.78 and $p = 2.95\times 10^{-220}$), *CDC25A* and *E2F2* (coefficient = 0.77 and $p = 3.99\times 10^{-217}$), and *PDGFRA* and *TGFBR2* (coefficient = 0.74 and $p = 3.66\times 10^{-186}$).

3.1.4. Mutation Attributes. $\chi2$ test was conducted to evaluate whether there was a significant difference in the mutation of a specific gene between two overall survival statuses. The results revealed that, among all the mutations, a total of 12 genes exhibited a statistically significant difference in mutation frequencies between the two statuses, shown in table 2.

Table 2: $\chi2$ and p values of mutation attributes with significantly different distributions across survival statuses.

Genes	$\chi2$	P value
MUC16	4.24	3.94×10^{-2}
GATA3	15.21	9.63×10^{-5}
DNAH2	5.75	1.65×10^{-2}
CBFB	6.02	1.41×10^{-2}
AKT1	4.91	2.67×10^{-2}
ARID1B	4.70	3.02×10^{-2}
PIK3R1	3.91	4.79×10^{-2}
BRCA2	4.44	3.52×10^{-2}
CACNA2D3	3.85	4.98×10^{-2}
FAM20C	3.85	4.98×10^{-2}
SMARCC1	3.94	4.71×10^{-2}
PALLD	6.24	1.25×10^{-2}

3.2. Modeling

3.2.1. Clinical Attributes. When training with clinical attributes, SVC outperformed among four models regarding ROC-AUC (AUC=0.75, figure 10). For performance metrics, SVC had an accuracy of 0.69, precision of 0.64, recall of 0.59, and F1 score of 0.61. The top three attributes were "integrative cluster 5", "age at diagnosis", and "tumor stage 0" (figure 11). The performance metrics of all four models are shown in table 3 in detail, and visualized in figure 9. Overall, four models had similar performance with clinical features.

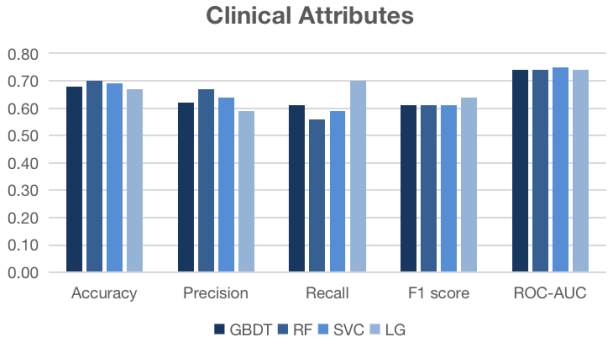

Figure 9. Histograms of performance metrics of four models trained with clinical attributes. The horizontal axe shows five metrics, including accuracy, precision, recall, and F1 score. Within each metric, GBDT, RF, SVC, and LG are presented from left to right. The horizontal axe presents the value of metrics.

DOI: 10.54254/2753-8818/40/20241235

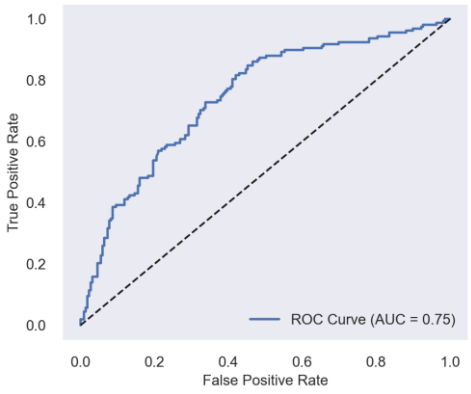

Figure 10. ROC curve of SVC trained with clinical attributes. The horizontal axis refers to the false positive rate, while the vertical axis refers to the true positive rate. The area under the curve is the value of AUC-ROC.

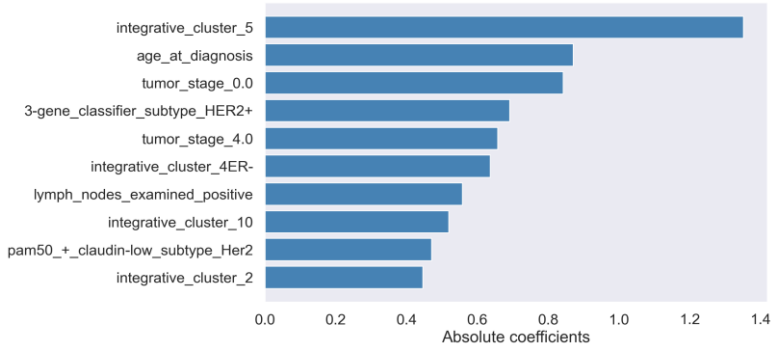

Figure 11. Bar graph of feature importance derived by SVC with clinical data. The horizontal axis presents the value of importance, while the vertical axis lists clinical attributes.

Table 3. Performance metrics of four models trained with clinical attributes.

Performance Metrics	GBDT	RF	SVC	LG
Accuracy	0.68	0.70	**0.69**	0.67
Precision	0.62	0.67	**0.64**	0.59
Recall	0.60	0.56	**0.59**	0.70
F1 score	0.61	0.61	**0.61**	0.64
ROC-AUC	0.74	0.74	**0.75**	0.74

3.2.2. Genetic Attributes. GBDT outperformed with genetic attributes regarding ROC-AUC (AUC=0.69, figure 13). For performance metrics, GBDT had an accuracy of 0.64, precision of 0.63, recall of 0.37, and F1 score of 0.47. The top three attributes were HSD17B11, JAK1, and STAT5A, (figure 14). The performance metrics of all four models are shown in table 4. in detail, and visualized in figure 12.

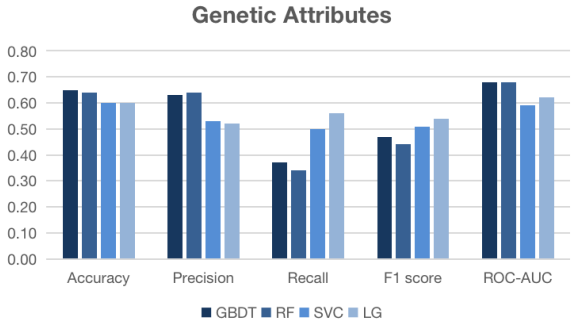

Figure 12. Histograms of performance metrics of four models trained with genetic attributes.

DOI: 10.54254/2753-8818/40/20241235

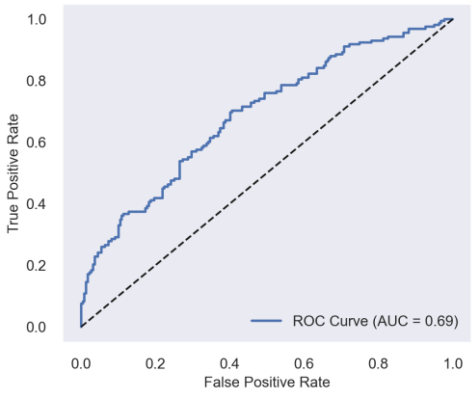

Figure 13. ROC curve of GBDT trained with genetic attributes.

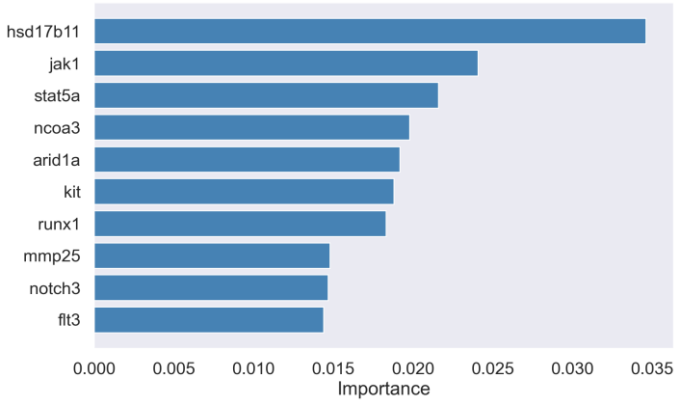

Figure 14. Bar graph of feature importance derived by GBDT with genetic data.

Table 4. Performance metrics of four models trained with genetic attributes.

Performance Metrics	GBDT	RF	SVC	LG
Accuracy	**0.64**	0.64	0.60	0.60
Precision	**0.63**	0.64	0.53	0.52
Recall	**0.37**	0.34	0.50	0.56
F1 score	**0.47**	0.44	0.51	0.54
ROC-AUC	**0.69**	0.68	0.59	0.62

3.2.3. Mutation Attributes. RF led the performance of four models trained with mutation attributes regarding ROC-AUC (AUC=0.58, figure 16). For performance metrics, RF had an accuracy of 0.58, precision of 0.50, recall of 0.41, and F1 score of 0.45. The top three mutations were PIK3CA, TP53, and AHNAK2 (figure 17). The performance metrics of all four models are shown in table 5 in detail, and visualized in figure 15.

DOI: 10.54254/2753-8818/40/20241235

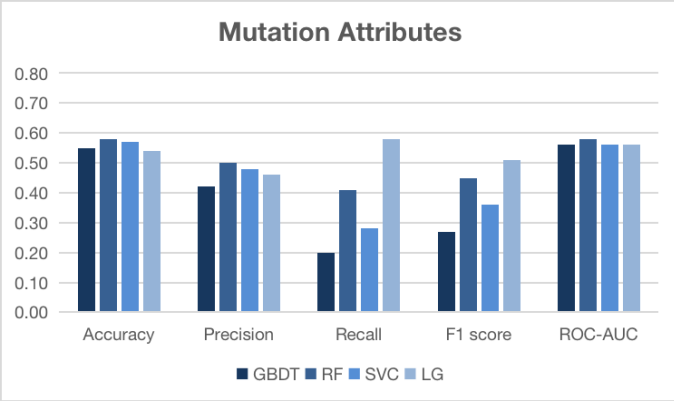

Figure 15. Histograms of performance metrics of four models trained with mutation attributes

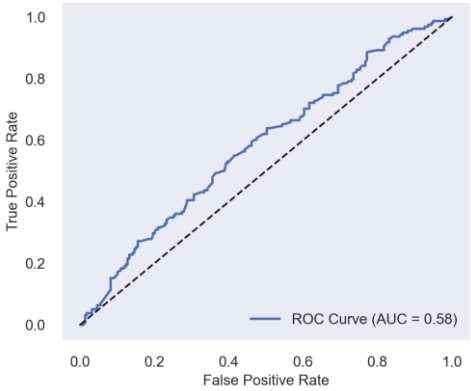

Figure 16. ROC curve of RF trained with mutation attributes.

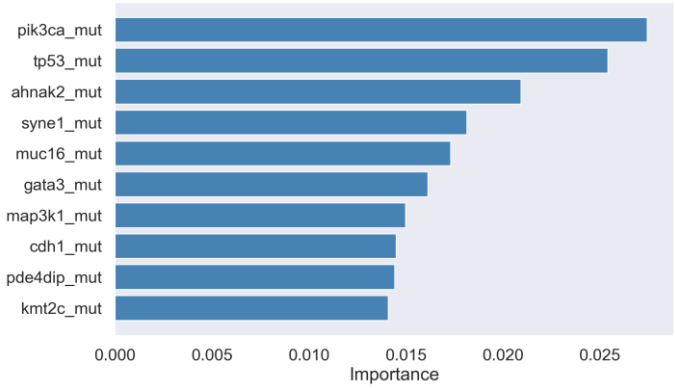

Figure 17. Bar graph of feature importance derived by RF with mutation data.

Table 5. Performance metrics of four models trained with mutation attributes.

Performance Metrics	GBDT	RF	SVC	LG
Accuracy	0.56	**0.58**	0.57	0.54
Precision	0.45	**0.50**	0.48	0.46
Recall	0.21	**0.41**	0.28	0.58
F1 score	0.29	**0.45**	0.36	0.51
ROC-AUC	0.55	**0.58**	0.56	0.56

3.2.4. Combined Attributes. GBDT ranked the top regarding ROC-AUC (AUC=0.75, figure 19) when trained with all attributes. For performance metrics, GBDT had an accuracy of 0.71, precision of 0.68, recall of 0.56, and F1 score of 0.61. The top three attributes were "age at diagnosis", "lymph nodes examined positive", and "Nottingham prognostic index" (figure 20). The performance metrics of all four models are shown in table 6 in detail, and visualized in figure 18. Overall, when trained with all attributes, GBDT and RF exhibited similar and higher performance, while SVC and LG showed close and relatively lower performance (figure 21). As for the significance of attributes, "age at diagnosis" and "lymph nodes examined positive" turned out to be important features for all four models, and interestingly, SVC and LG shared the same key features which had only a few overlaps with GBDT and RF (figure 22).

DOI: 10.54254/2753-8818/40/20241235

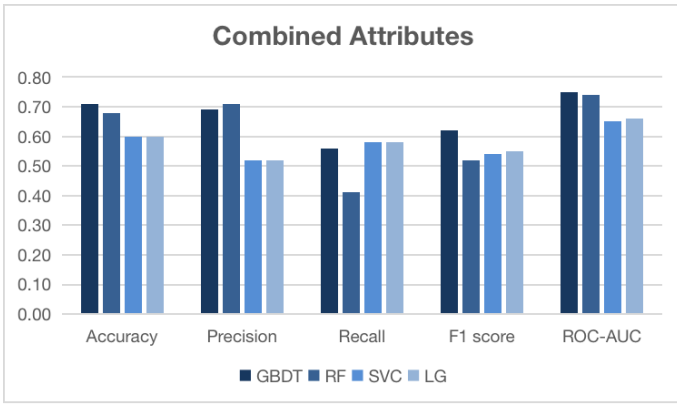

Figure 18. Histograms of performance metrics of four models trained with all attributes.

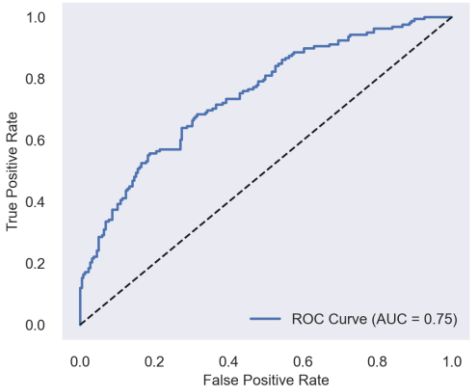

Figure 19. ROC curve of GBDT trained with all attributes.

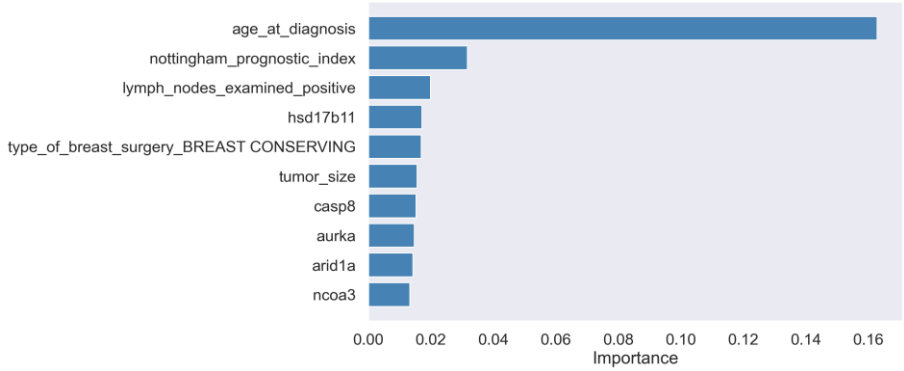

Figure 20. Bar graph of feature importance derived by GBDT with all data.

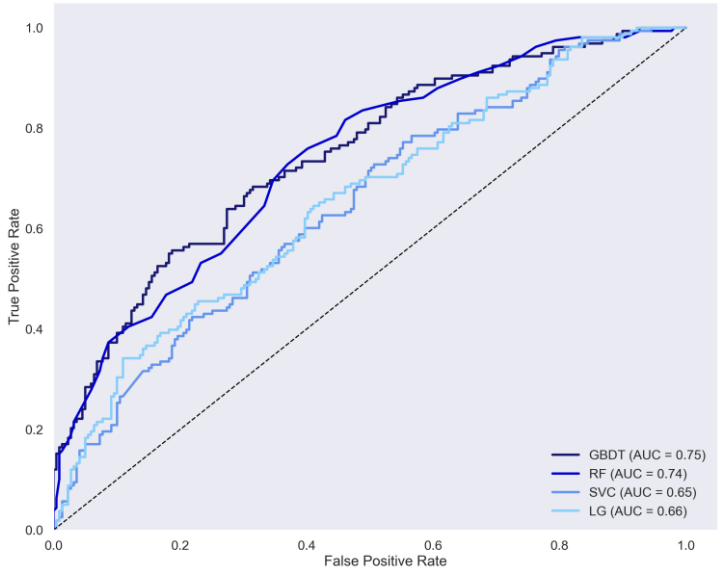

Figure 21. ROC curve of four models trained with all attributes. Ranging from the darkest to the lightest, four different degrees of blue represent GBDT, RF, SVC, and LG, respectively.

DOI: 10.54254/2753-8818/40/20241235

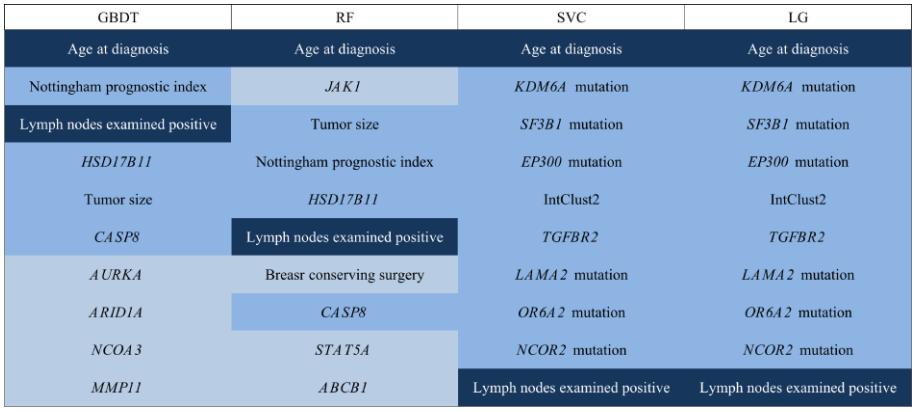

Figure 22. Top 10 features of four models trained with all attributes. The deepest blue refers to features that appeared in four models, the middle blue represents features that appeared in any two models, and the lightest blue that indicates features appeared in only one model.

Table 6. Performance metrics of four models trained with mutation attributes.

Performance Metrics	GBDT	RF	SVC	LG
Accuracy	**0.71**	0.68	0.60	0.60
Precision	**0.68**	0.71	0.52	0.52
Recall	**0.56**	0.41	0.58	0.58
F1 score	**0.61**	0.52	0.54	0.55
ROC-AUC	**0.75**	0.74	0.65	0.66

4. Discussion

In conclusion, the dataset was trained with four models, including LG, SVC, RF, and GBDT, and important features were identified for prognosis among clinical, genetic, and mutation attributes. Overall, when all features were combined for training, GBDT had the best performance, followed by RF, LG, and SVC. In addition, the top three attributes for GBDT with combined features were “age at diagnosis”, “lymph nodes examined positive”, and “Nottingham prognostic index”.

Each model involved in training has its pros and cons of predicting the outcome of breast cancer. LG performs well with linearly separable datasets. Coefficients in its weighted sum can be interpreted as feature importance, giving insights into the selection of important features (such as “age at diagnosis”, “Nottingham prognostic index”, etc.) relevant to the prognosis of breast cancer. A positive coefficient indicates the feature is more important for the category labeled as 1, which refers to survival, while a negative coefficient indicates the importance for the category labeled as 0, which refers to death. However, LG is not suitable for unbalanced and nonlinear datasets and is sensitive to outliers, which are high demands on the original data. In addition, LG cannot learn non-linear relationships of variables. However, as for our data, the “survived” group and the “died” group accounted for 0.42% and 0.58% of the total respectively, which indicates that the dataset was unbalanced. Additionally, the dataset consisted of 687 features, including both numeric and categorical attributes, meaning that our dataset had high dimensionality and non-linearity. However, SVC, unlike LG, can handle more types of datasets, as SVC has flexible kernels including linear, polynomial, sigmoid, and radial basis functions. SVC is effective in high-dimensional space. Still, SVC also has obvious drawbacks. First of all, the computational cost of SVC is significantly higher than that of other models. On top of that, SVC has low interpretability, as SVC only finds the hyperplane that maximizes the margin of data in the space. SVC also has a high risk of overfitting. Similar to SVC, RF is a robust algorithm that is suitable for data with high dimensionality, outliers, noises, and nonlinearity. But RF has advantages over SVC in that RF reduces overfitting by involving different decision trees, and offers an interpretation of feature importance. However,

DOI: 10.54254/2753-8818/40/20241235

overfitting would still happen when the number and depth of trees are too high. For GBDT, it exhibits superior performance compared with RF regarding accuracy. GBDT combines multiple weak learners to make a strong learner by reducing residuals in the training process. GBDT is extremely powerful but takes a longer time to train as GBDT requires proper hyperparameter tuning.

According to the results of statistics, attributes with significantly different distributions across the overall status were selected to train four models again, which turned out that the performance of LG and SVC was enhanced. For genetic features, the AUC of LG improved from 0.62 to 0.66, while the AUC of SVC increased from 0.59 to 0.66. With combined features, the AUC of LG rose from 0.66 to 0.71, while the AUC of SVC increased from 0.65 to 0.70. However, the performance of GBDT and RF remained almost unchanged. At last, PCA was conducted and the result is shown in figure 23, which indicates that 150 PCs derived from 687 features are capable enough to explain over 80% variance of the dataset.

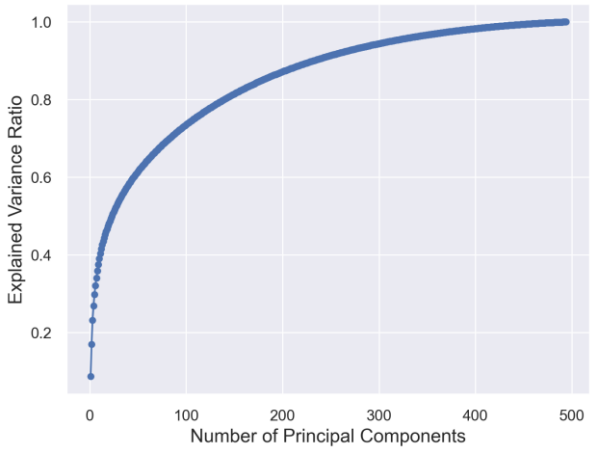

Figure 23. Correlation between the number of principal components and explained variance ratio. The horizontal axis represents the number of PCs, while the vertical axis represents the variance ratio explained by the corresponding number of PCs.

Many studies have employed machine learning and deep learning approaches for breast cancer survival prediction using clinical data. “Cancer stage classification”, “tumor size”, “number of total axillary lymph nodes removed”, “number of positive lymph nodes”, “types of primary treatment”, and “methods of diagnosis” were identified as the most important features for the outcome prediction [15,22]. These previous findings are consistent with the present study that, when training the models only with clinical data, “age at diagnosis”, “tumor size”, and “lymph nodes examined positive” were also identified as the most relevant features. Additionally, “Nottingham prognostic index”, “mutation count”, “intrinsic subtypes of Luminal A and HER2-enriched”, “integrative clusters of IntClust 5”, and “type of breast surgery” emerged among the top ten variables.

While most of the previous studies on survival prediction were limited to clinical features, genetic expression, and mutation variables were also taken into consideration in this study. Among the genetic features, the expressions of *HSD17B11*, *JAK1*, *STAT5A*, *NCOA3,* and *KIT* were identified as the most influential factors for the outcome prediction. In the case of mutation variables, mutations in *GATA3*, *TP53*, *MUC16*, *LAMA2*, and *PIK3CA* stand out among other genes.

Those results have supportive evidence in other studies. The importance of “age at diagnosis”, “tumor size”, “mutation count”, and “lymph nodes examined positive” in survival prediction has been verified in previous studies [15,22]. In a project conducted by the Molecular Taxonomy of Breast Cancer International Consortium [23], IntClust 5 (weight = -1.01, LG of the present study) stands out with the extremely poor outcome among all integrative clusters, providing a reasonable basis for IntClust 5 to be one of the most important features for outcome prediction. Intrinsic subtypes of Luminal A (weight = 0.26) and HER2-enriched (weight = 0.57) both have a relatively good prognosis, as Luminal A is relatively low grade and HER2-enriched tumor responds aggressively to anti-HER2-targeted therapy [8]. *HSD17B11* (weight = 0.61) is responsible for the expression of hydroxysteroid 17-beta dehydrogenase 11, whose high level is associated with improved relapse-free survival in breast cancer [24]. A low expression level of *JAK1* (weight = 0.35) is associated with a low survival rate, as the

DOI: 10.54254/2753-8818/40/20241235

expression of Janus Kinase 1 is negatively correlated with tumor size, number of lymph nodes, and stage status, and Janus Kinase 1 is positively correlated with infiltrating levels of immune cells [25]. For *STAT5A* (weight = 0.20), it expresses signal transducer and activator of transcription 5A (Stat5a), which is generally downregulated in breast cancer [26]. It was demonstrated that the low expression of Stat5a protein is often correlated with tumor progression and unfavorable clinical outcomes [27]. As for *NCOA3* (weight = -0.20), it is responsible for the expression of nuclear receptor coactivator 3 (NCoA3), which is often involved in cell resistance to chemotherapy, leading to poor prognosis [28]. Additionally, *NCOA3* is upregulated in breast cancer-associated adipocytes and has an impact on inflammation [29]. *KIT* (weight = 0.05) encodes proto-oncogene receptor tyrosine kinase which plays a crucial role in cell proliferation. A previous study revealed that the loss of *KIT* expression was correlated with malignant transformation of breast epithelium [30]. Mutations in breast cancers are infrequent. Among all types of mutation, *PIK3CA* (weight = -0.31) and *TP53* (weight = -0.22) dominate, occurring at incidences of 40.1% and 35.4% respectively [10]. As for *MUC16*, it is commonly expressed in ovarian cancer and can be identified in other epithelial-origin tumors including gastric cancer and breast cancer [31,32]. It was demonstrated that *MUC16* regulates tumor growth and the slicing of *MUC16* could lead to a decrease in cell adhesion, migration, and invasion [32]. Furthermore, previous evidence also noted that *LAMA2* was poorly expressed in breast cancer tissues and associated with a low survival rate [33].

For the next step, Weighted Gene Co-expression Network Analysis (WGCNA) could be applied to simplify the interpretation of multiple genes within the dataset and gain deeper insights into the gene correlations [34]. By studying the co-expression network, the drawback of focusing only on individual genes in our study could be addressed [35], which would help to further identify important genes for prognosis. Furthermore, deep learning methods such as Multi-layer Perceptron (MLP) can be employed for superior performance in outcome prediction, as deep learning models possess the capability to learn with minimal human intervention or preconceived assumption, providing deep learning models with an edge over machine learning [15].

5. Conclusion

In conclusion, this study analyzed a dataset encompassing clinical, genetic, mutation features, and survival status of 1882 breast cancer patients to gain insights into prognostic factors. Multiple statistical tests and four machine learning models, including LG, SVC, RF, and GBDT, were employed for the in-depth study. Key clinical attributes, such as “age at diagnosis”, “lymph nodes examined positive”, and “Nottingham prognostic index”, along with gene expressions including *HSD17B11*, *JAK1*, and *STAT5A*, as well as mutations in *GATA3*, *TP53*, and *MUC16*, were identified as top relative attributes. Furthermore, regarding performance in predicting breast cancer prognosis, GBDT demonstrated the highest AUC-ROC compared to LG, SVC, and RF.

References

[1] Smolarz B, Nowak AZ and Romanowicz H. 2022. Breast Cancer-Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). *Cancers (Basel)*. 14(10):2569.

[2] Siegel RL, Miller KD, Wagle NS and Jemal A. 2023. Cancer statistics, 2023. *CA Cancer J Clin*. 73(1):17–48.

[3] Mattiuzzi C and Lippi G. 2019. Current Cancer Epidemiology. *J Epidemiol Glob Health*. 9(4):217–22.

[4] Henry NL and Cannon-Albright LA. 2019. Breast cancer histologic subtypes show excess familial clustering. *Cancer*. 125(18):3131–8.

[5] Rakha EA et al. 2010. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res*. 12(4):207.

[6] Prat A et al. 2015. Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast*. 24 Suppl 2:S26-35.

[7] Makki J. 2015. Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. *Clin Med Insights Pathol*. 8:23–31.

[8] Tsang JYS and Tse GM. 2020. Molecular Classification of Breast Cancer. *Adv Anat Pathol*. 27(1):27–35.
[9] Russnes HG, Lingjærde OC, Børresen-Dale AL and Caldas C. 2017. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *Am J Pathol*. 187(10):2152–62.
[10] Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. *Nature*. 490(7418):61–70.
[11] Kuchenbaecker KB et al. 2017. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA*. 317(23):2402–16.
[12] Kashyap D et al. 2022. Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures. *Biomed Res Int*. 2022:9605439.
[13] Dey S, Boffetta P, Mathews A, Brennan P, Soliman A and Mathew A. 2009. Risk factors according to estrogen receptor status of breast cancer patients in Trivandrum, South India. *Int J Cancer*. 125(7):1663–70.
[14] Li C et al. 2022. Machine learning predicts the prognosis of breast cancer patients with initial bone metastases. *Front Public Health*. 10:1003976.
[15] Kalafi EY, Nor N a. M, Taib NA, Ganggayah MD, Town C and Dhillon SK. 2019. Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data. *Folia Biol (Praha)*. 65(5–6):212–20.
[16] Nguyen QTN et al. 2023. Machine learning approaches for predicting 5-year breast cancer survival: A multicenter study. *Cancer Sci*. 114(10):4063–72.
[17] Montazeri M, Montazeri M, Montazeri M and Beigzadeh A. 2016. Machine learning models in breast cancer survival prediction. *Technology and Health Care*. 24(1):31–42.
[18] Zhang X et al. 2019. Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int J Med Inform*. 132:103985.
[19] Lee AHS and Ellis IO. 2008. The Nottingham Prognostic Index for Invasive Carcinoma of the Breast. *Pathol. Oncol. Res.* 14(2):113–5.
[20] Lever J, Krzywinski M and Altman N. 2017. Points of significance: Principal component analysis. *Nature methods*. 14(7):641–3.
[21] Ringnér M. 2008. What is principal component analysis? *Nature biotechnology*. 26(3):303–4.
[22] Ganggayah MD, Taib NA, Har YC, Lio P and Dhillon SK. 2019. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak*. 19(1):48.
[23] Rueda OM et al. 2019. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature*. 567(7748):399–404.
[24] Corbet AK et al. 2023. G0S2 promotes antiestrogenic and pro-migratory responses in ER+ and ER- breast cancer cells. *Transl Oncol*. 33:101676.
[25] Chen B, Lai J, Dai D, Chen R, Li X and Liao N. 2019. JAK1 as a prognostic marker and its correlation with immune infiltrates in breast cancer. *Aging (Albany NY)*. 11(23):11124–35.
[26] Mukhopadhyay UK et al. 2016. Dataset of STAT5A status in breast cancer. *Data Brief*. 7:490–2.
[27] Peck AR et al. 2012. Low levels of Stat5a protein in breast cancer are associated with tumor progression and unfavorable clinical outcomes. *Breast Cancer Res*. 14(5):R130.
[28] Rubio MF, Lira MC, Rosa FD, Sambresqui AD, Salazar Güemes MC and Costas MA. 2017. RAC3 influences the chemoresistance of colon cancer cells through autophagy and apoptosis inhibition. *Cancer Cell Int*. 17(1):111.
[29] Lira MC et al. 2023. NCoA3 upregulation in breast cancer-associated adipocytes elicits an inflammatory profile. *Oncol Rep*. 49(5):105.
[30] Janostiak R, Vyas M, Cicek AF, Wajapeyee N, and Harigopal M. 2018. Loss of c-KIT expression in breast cancer correlates with malignant transformation of breast epithelium and is mediated by KIT gene promoter DNA hypermethylation. *Exp. Mol. Pathol.* 105(1):41–9.
[31] Li X, Pasche B, Zhang W and Chen K. 2018. Association of MUC16 Mutation With Tumor Mutation Load and Outcomes in Patients With Gastric Cancer. *JAMA Oncol*. 4(12):1691–8.

DOI: 10.54254/2753-8818/40/20241235

[32] Reinartz S, Failer S, Schuell T and Wagner U. 2012. CA125 (MUC16) gene silencing suppresses growth properties of ovarian and breast cancer cells. *Eur J Cancer*. 48(10):1558–69.
[33] Li S et al. 2023. Epigenetic regulation of LINC01270 in breast cancer progression by mediating LAMA2 promoter methylation and MAPK signaling pathway. *Cell Biol Toxicol*. 39(4):1359–75.
[34] Yin X et al. 2020. Identification of key modules and genes associated with breast cancer prognosis using WGCNA and ceRNA network analysis. *Aging (Albany NY)*. 13(2):2519–38.
[35] Giulietti M, Occhipinti G, Principato G and Piva F. 2017. Identification of candidate miRNA biomarkers for pancreatic ductal adenocarcinoma by weighted gene co-expression network analysis. *Cell Oncol (Dordr)*. 40(2):181–92.