Scientific Research Publishing

Advances in Breast Cancer Research, 2023, 12, 163-185
https://www.scirp.org/journal/abcr
ISSN Online: 2168-1597
ISSN Print: 2168-1589

Decipher Clinical and Genetic Underpins of Breast Cancer Survival with Machine Learning Methods

Zhengkai Zhuang

College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
Email: u202016408@hust.edu.cn

How to cite this paper: Zhuang, Z.K. (2023) Decipher Clinical and Genetic Underpins of Breast Cancer Survival with Machine Learning Methods. *Advances in Breast Cancer Research*, **12**, 163-185.
https://doi.org/10.4236/abcr.2023.124013

Received: September 11, 2023
Accepted: October 23, 2023
Published: October 26, 2023

Abstract

Breast cancer is one of the most common cancers among women in the world, with more than two million new cases of breast cancer every year. This disease is associated with numerous clinical and genetic characteristics. In recent years, machine learning technology has been increasingly applied to the medical field, including predicting the risk of malignant tumors such as breast cancer. Based on clinical and targeted sequencing data of 1980 primary breast cancer samples, this article aimed to analyze these data and predict living conditions after breast cancer. After data engineering, feature selection, and comparison of machine learning methods, the light gradient boosting machine model was found the best with hyperparameter tuning (precision = 0.818, recall = 0.816, f1 score = 0.817, roc-auc = 0.867). And the top 5 determinants were clinical features age at diagnosis, Nottingham Prognostic Index, cohort and genetic features rheb, nr3c1. The study shed light on rational allocation of medical resources and provided insights to early prevention, diagnosis and treatment of breast cancer with the identified risk clinical and genetic factors.

Keywords

Machine Learning, Breast Cancer Prediction, Data Analysis, Feature Importance Comparison

1. Introduction

As the most common cancer worldwide and the leading cause of cancer deaths among women, breast cancer brings serious health and social difficulties to society around the world. Breast cancer is a phenomenon in which mammary epithelial cells proliferate out of control under the action of various carcinogens.

According to the data from WHO, there are more than 2.5 million cases each year and the incidence and number of lives lost to breast cancer is increasing [1].

The etiology of breast cancer has not been completely clarified, but some factors related to its increased risk have been identified. For example, family history of breast can be a factor of increasing the risk of breast cancer, including many genetic mutations in key genes. There is evidence of many factors that contribute to breast cancer risk, including the reproductive, lifestyle, environmental factors, etc. [2]. And they can be reflected by features clinically.

In addition, the diagnosis methods of breast cancer include physical examination, breast ultrasound imaging, mammography and pathological biopsy [3] [4]. Regular physical examination is likely to find breast cancer in an early stage and it is one of the simplest and most economical methods of judgment methods. Ultrasound imaging can show the location, shape and texture of breast mass, and judge whether it is benign or malignant. Mammography is the most effective modality in detection and diagnosis of breast cancer [5]. For pathological biopsy, it is one of the most reliable diagnostic methods for breast cancer, which can be diagnosed by means of histological features, immunohistochemical markers and gene detection.

The treatment of breast cancer mainly includes surgical resection, radiotherapy, chemotherapy, endocrine therapy and targeted therapy [6] [7] [8].

In recent years, with the development of machine learning methods, more and more related algorithms have been applied in medical and clinical related fields, mainly including disease prediction, auxiliary diagnosis, and so on. A machine learning algorithm helps a lot to make decisions and to perform diagnosis from the data collected by the medical field. In the present study, many models were applied to breast cancer prediction [9]. For example, various normal machine learning classification methods like Logistic Regression, K Neighbors Classifier, Naïve Bayes, support vector machine (SVM), Decision Tree Classifier and many ensemble techniques including Random Forest, Light Gradient Boosting Machine, Gradient Boosting Classifier have been well applied [10] [11] [12]. These models were compared with each other and selected, important features were found hoping to offer help in the aspect of diagnosis and treatment (top three models: Light Gradient Boosting Machine, Gradient Boosting Classifier, K Neighbors Classifier, three most important features: age at diagnosis, cohort, Nottingham Prognostic Index). The source of the data comes from The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database and is downloaded from cBioPortal [13].

2. Methods

2.1. Feature Description

2.1.1. Numerical Clinical Features

Numerical features included age, lymph nodes examined positive, mutation count, Nottingham Prognostic Index, overall survival months and tumor size. Age at diagnosis means the age of the patient at diagnosis time. Lymph nodes examined positive refer to the number of the positive lymph nodes after histopathological exam. Mutation count means the number of genes which have mutations. Nottingham Prognostic Index is used as an index related to concerning grade, tumor size, and lymph node status, each weighted according to regression coefficients of a Cox proportional hazard analysis [14]. Overall survival months means the duration from the time of the intervention to death. Tumor size shows the tumor size measured by imaging techniques.

2.1.2. Categorical Clinical Features

Categorical features included 22 features. Cancer type is the type of cancer which can be separated into 2 types: Breast Cancer and Breast Sarcoma. Type of breast surgery refers to breast cancer surgery type. And the main treatment method of breast cancer continues to be surgical: breast-conserving surgery, typically with adjuvant radiation, or mastectomy. They are the main two types [15]. “Mastectomy” refers to a surgery to remove all breast tissue from breast as a way to treat or prevent breast cancer. “Breast conserving” refers to a surgery where only the part of the breast that has cancer is removed. Cancer type detailed is the more detailed cancer types. It can be separated into: Breast Invasive Ductal Carcinoma, Breast Mixed Ductal and Lobular Carcinoma Breast Invasive Lobular Carcinoma, Breast Invasive Mixed Mucinous Carcinoma and Metaplastic Breast Cancer. Cellularity refers to the amount of tumor cells in the specimen and their arrangement into clusters. According to the amount level, we can separate it into three groups: low, moderate and high. Chemotherapy is whether the patient had chemotherapy as a treatment or not. “1” refers to yes and “0” refers to no. For Pam50 + claudin-low subtype, Pam 50 is a tumor profiling test that helps show whether some estrogen receptor-positive (ER-positive), HER2-negative breast cancers are likely to metastasize (when breast cancer spreads to other organs). According to Perou *et al.*, there are initially four breast cancer intrinsic subtypes including basal-like, HER2-enriched, luminal and normal-like, which can display gene expression patterns. Subsequent studies have led to the sub-stratification of luminal breast cancers into luminal A and luminal B, and shown that this classification system is of prognostic significance [16]. The classified information was shown in **Table 1**.

Cohort refers to a group of subjects who share a defining characteristic (It takes a value from 1 to 5). Er status measured by ihc is by using the IHC method, the tumour was considered oestrogen receptor (ER) positive if ≥10% of the neoplastic cells showed nuclear staining, otherwise it will be negative [17]. Er status refers to the estrogen receptor-alpha (ER-α), the binary ER status is predictive for treatment response and prognostic for outcome [18]. Neoplasm histologic grade is determined by pathology by looking the nature of the cells, to judge they look aggressive or not (It takes a value from 1 to 3). Her2 status measured by snp6 is used to assess if the cancer positive for HER2 or not by

Table 1. Subtypes of breast cancer.

Subtypes of breast cancer	Classification index
Luminal A	ER and/or PR positive HER2 negative Ki-67 < 14%
Luminal B (HER2 negative)	ER and/or PR positive HER2 negative Ki-67 ≥ 14%
Luminal B (HER2 positive)	ER and/or PR positive HER2 overexpressed or amplified Any Ki-67
HER2-enriched	ER and PR absent HER2 overexpressed or amplified
Basal-like	ER and PR absent HER2 negative

ER: oestrogen receptor; PR: progesterone receptor; HER2: human epidermal growth factor receptor 2; Ki-67: a kind of proliferation marker associated with the expression of ER-associated genes.

using advance molecular techniques (type of next generation sequencing). Her2 status is whether the cancer is positive or negative for HER2. Tumor other histologic subtype is the type of the cancer based on microscopic examination of the cancer tissue (It takes a value of "Ductal/NST", "Mixed", "Lobular", "Tubular/cribriform", "Mucinous", "Medullary", "Other", "Metaplastic"). Hormone therapy refers to whether or not the patient had hormonal as a treatment (1-yes/0-no). Inferred menopausal state is whether the patient postmenopausal or not (post/pre). Integrative cluster is the molecular subtype of the cancer based on some gene expression (it takes a value from "4ER+", "3", "9", "7", "4ER−", "5", "8", "10", "1", "2", "6"). Primary tumor laterality is whether the right breast or the left breast is involved. For Oncotree code, the OncoTree is an open-source ontology that was developed at Memorial Sloan Kettering Cancer Center (MSK) for standardizing cancer type diagnosis from a clinical perspective by assigning each diagnosis a unique OncoTree code. Pr status refers to Cancer cells are positive or negative for progesterone receptors. Radio therapy is whether or not the patient had radio as a treatment (yes-1/no-0). 3-gene classifier subtype is the three Gene classifier subtype. It takes a value from "ER−/HER2−", "ER+/HER2− High Prolif", "ER+/HER2− Low Prolif", "HER2+". Tumor stage refers to the stage of the cancer based on the involvement of surrounding structures, lymph nodes and distant spread.

2.1.3. Genetic Features

This part of data included mRNA levels Z-score for 331 genes collected from present research. They were proved to have relationship with breast cancer [13]. mRNA expression data were used to calculate the relative expression of an individual gene and tumor to the gene's expression distribution. For all samples in the returned value indicated the number of standard deviations away from the mean of expression in the reference population (Z-score). This measure was useful to determine whether a gene is up- or down-regulated relative to the normal samples or all other tumor samples. The formula was

$$Z = \frac{\text{expression in tumor sample} - \text{mean expression in reference sample}}{\text{standard deviation of expression in reference sample}}.$$

2.1.4. Mutation Features

This part of data included mutation information for 175 genes which were related to breast cancer [13]. "1" means there are mutations on the gene while "0" means there are no mutations.

2.1.5. Target Feature

Overall survival refers to whether the patient is alive (1) or dead (0).

2.2. Statistics

To further analyze the data, Student's t-test and χ^2 tests were mainly used based on target feature. Using survival or non-survival as classification criteria, all data were split into two groups. For numerical features, two-sample t-test was used to justify whether the difference between the average of two samples and their respective populations was significant or not. The formula is

$$t = \frac{\underline{x_1} - \underline{x_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

($\underline{x_1}$ and $\underline{x_2}$ are the average value of survival and non-survival group, S_1^2 and S_2^2 are the variances of the two groups, and n_1 and n_2 are the sample sizes of the two groups). P value is the probability of making a mistake by accepting the hypothesis that there is a difference between average values of the two groups. Normally, if P < 0.05, the feature tested between the two groups was normally thought significantly different. For categorical features, the independence test χ^2 analysis was used to verify whether the paired observation groups extracted from two variables were independent of each other. The formula is

$$\chi^2 = \sum \frac{(A-T)^2}{T}$$

(A is the observation value, and T is the theoretical value). The same as t-test on numerical features, P value is the probability of making a mistake by accepting the hypothesis that there is a difference between average values of the two groups. Normally, if P < 0.05, the features tested between the two groups were considered significantly different.

2.3. Machine Learning Method and Model Performance Evaluation Parameters

To determine the best model to fit the data, many machine learning methods including Light Gradient Boosting Machine, Gradient Boosting Classifier, K Neighbors Classifier, Ada Boost Classifier, Logistic Regression, Random Forest Classifier, SVM (Linear Kernel), Extra Trees Classifier, Decision Tree Classifier, Ridge Classifier, Linear Discriminant Analysis, Dummy Classifier, Quadratic Discriminant Analysis and Naive Bayes were used to train the data. To select the best method from these machine learning models, accuracy, precision, recall, f1_score and area under the receiver operating characteristic curve (AUC-ROC) were used as model performance evaluation parameters (In the following part, TP refers to the size of positive samples predicted by the model as positive, TN refers to the size of negative samples predicted by the model as negative, FP refers to the size of negative samples predicted by the model as positive, FN refers to the size of positive samples predicted by the model as negative). **Accuracy** = $\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$, which is the most common evaluation index used in evaluation. The accuracy of training data and test data was tested. **Precision** = $\frac{\text{TP}}{\text{TP}+\text{FP}}$, it is the index to examine the percent of true positive in all samples which is predicted positive. It was used to exam the test samples. **Recall** = $\frac{\text{TP}}{\text{P}}$, it is the index to examine the percentage of true positives in all samples which are really positive. **F1** = $\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$, which is the weighted harmonic average of precision and recall. AUC-ROC curve is a performance measure of classification problems under different threshold settings. ROC means the probability curve, and AUC means the degree to which positive and negative categories can be correctly classified. It tells the model to what extent it can be classified. ROC curve is plotted by True Positive Rate (TPR) and False Positive Rate (FPR), where TPR is the y axis and FPR is the x axis.

3. Results

3.1. Analysis on Feature Based on Target Feature

3.1.1. Student's T-Tests on Numerical Feature

Through two-sample t-tests on numerical feature based on survival and non-survival group (There are 801 samples in survival group and 1103 samples in non-survival group), all 6 numerical features were significantly different between survival and non-survival group. Age at diagnosis was statistically significantly lower in the survival group (survival = 56.46 ± 11.37 yr, non-survival = 64.44 ± 13.04 yr, P = 6.61×10^{-42}), lymph nodes examined was statistically significantly lower in the survival group (survival = 1.21 ± 2.72, non-survival = 2.57 ± 4.75, P = 5.11×10^{-13}), mutation count was significantly different between survival and non-survival group (survival = 5.32 ± 3.34, non-survival = 5.96 ± 4.48, P = 7.94×10^{-4}), Nottingham Prognostic Index was statistically significantly lower in the survival group (survival = 3.85 ± 1.07, non-survival = 4.17 ± 1.18, P = 1.48 × 10^{-9}), overall survival months was statistically significantly lower in the non-survival group (survival = 159.55 ± 71.65 months, non-survival = 100.12 ± 69.57 months, P = 4.11 × 10^{-68}), and tumor size was statistically significantly lower in the survival group (survival = 23.32 ± 13.05, non-survival = 28.36 ± 16.19, P = 7.06 × 10^{-13}).

3.1.2. χ^2 Analysis on Categorical Feature

χ^2 analysis was also used on categorical feature based on survival and non-survival groups. Nine out of all 22 categorical features were significantly different between survival and non-survival groups, including type of breast surgery, cancer type detailed, chemotherapy Pam50 + claudin-low subtype, and so on. Type of breast surgery is statistically different between survival and non-survival groups (P = 4.24 × 10^{-16}). Cancer type is not statistically different between survival and non-survival groups (P = 0.87). Cancer type detailed is statistically different between survival and non-survival groups (P = 0.04). Cellularity is not statistically different between survival and non-survival groups (P = 0.32). Chemotherapy is not statistically different between survival and non-survival groups (P = 0.05). Pam50 + claudin-low subtype is statistically different between survival and non-survival groups P = 2.57 × 10^{-3}). Cohort is statistically different between survival and non-survival groups (P = 0.03). Er status measured by ihc is not statistically different between survival and non-survival groups (P = 0.38). Er status is not statistically different between survival and non-survival groups (P = 0.42). Neoplasm histologic grade is statistically different between survival and non-survival groups (P = 0.02). Her2 status measured by snp6 is not statistically different between survival and non-survival groups (P = 0.08). Her2 status is not statistically different between survival and non-survival groups (P = 0.17). Tumor other histologic subtype is not statistically different between survival and non-survival groups (P = 0.89). Hormone therapy is not statistically different between survival and non-survival groups (P = 0.20). Inferred menopausal state is statistically different between survival and non-survival groups (P = 1.35 × 10^{-13}). Integrative cluster is not statistically different between survival and non-survival groups (P = 0.11). Primary tumor laterality is not statistically different between survival and non-survival groups (P = 0.06). Oncotree code is statistically different between survival and non-survival groups (P = 0.04). Pr status is not statistically different between survival and non-survival groups (P = 0.35). Radio therapy is statistically different between survival and non-survival groups (P = 1.28 × 10^{-6}). 3-gene classifier subtype is statistically different between survival and non-survival groups (P = 1.86 × 10^{-6}). Tumor stage is not statistically different between survival and non-survival groups (P = 0.76).

3.2. Data Preprocessing

3.2.1. Missing Value Removal

For all features, whether it has missing value was detected, and the percentage of missing value was calculated. The information was shown in **Table 2**. According to the percentage, features were sorted in descending order.

All percent of missing value < 30%, the samples which include missing value were removed in the process of modeling.

3.2.2. Single Clinical Feature Analyses

For all clinical features, they were split into two groups: numerical features and categorical features.

All numerical features were selected out and checked, and their basic information was shown in Table 3.

For every numerical feature, its histogram and boxplot were drawn to see its distribution and frequency (Figure 1 and Figure 2).

All categorical clinical features were selected out. For every categorical feature, the numbers of different kinds in the categorical features were shown (Figures 3-6).

Table 2. Missing values.

Feature	Missing value count	Missing value percent
Tumor_stage	501	26.31%
3-gene_classifier_subtype	106	10.71%
Primary_tumor_laterality	72	5.57%
Neoplasm_histologic_grade	54	3.78%
Cellularity	45	2.84%
Mutation_count	30	2.36%
Er_status_measured_by_ihc	22	1.58%
Type_of_breast_surgery	20	1.16%
Tumor_size	15	1.05%
Tumor_other_histologic_subtype	15	0.79%
Cancer_type_detailed	15	0.79%
Oncotree_code	15	0.79%
Death_from_cancer	1	0.53%

Table 3. Basic information of numerical feature.

Feature	Count	Mean	Std	Min	Max
Age_at_diagnosis	1904	61.09	12.98	21.93	96.29
Lymph_nodes_examined_positive	1904	2.00	4.08	0	45
Mutation_count	1859	5.70	4.06	1	80
Nottingham_prognostic_index	1904	4.03	1.14	1	6.36
Overall_survival_months	1904	125.12	76.33	0	355.20
Tumor_size	1884	26.24	15.16	1	182

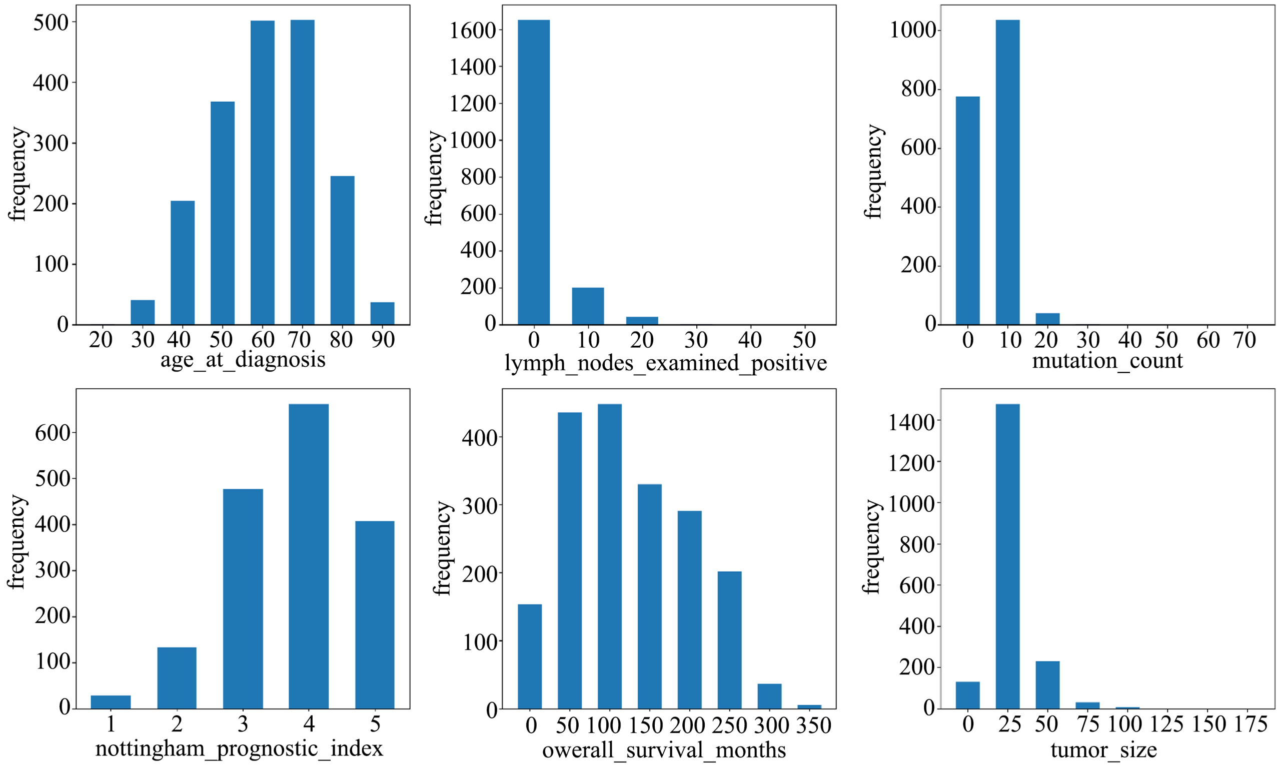

Figure 1. Histogram of numerical clinical features.

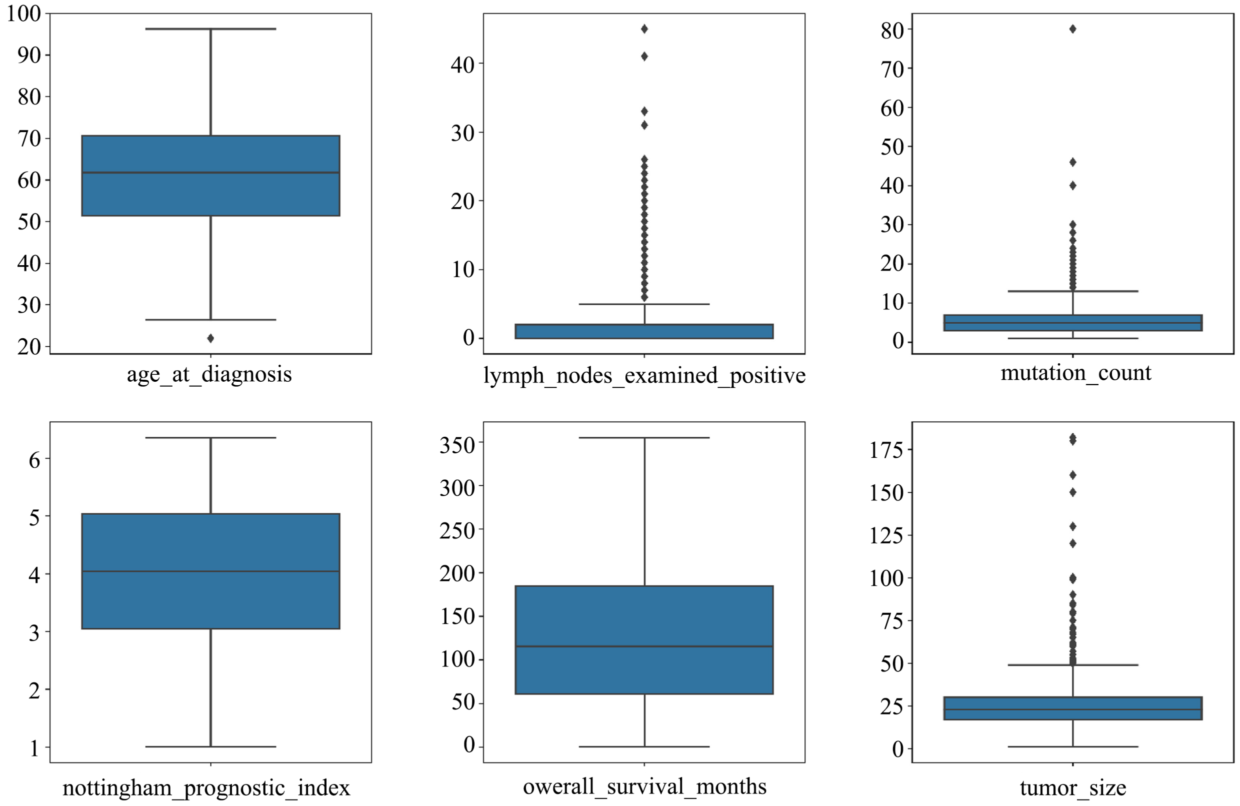

Figure 2. Boxplot of numerical clinical features. Upper bound (Q3 + 1.5IQR, IQR = Q3 − Q1), 75% quartile Q3, Median, 25% quartile Q1, Lower bound (Q1 − 1.5IQR). Outliers: outside of upper or lower bound.

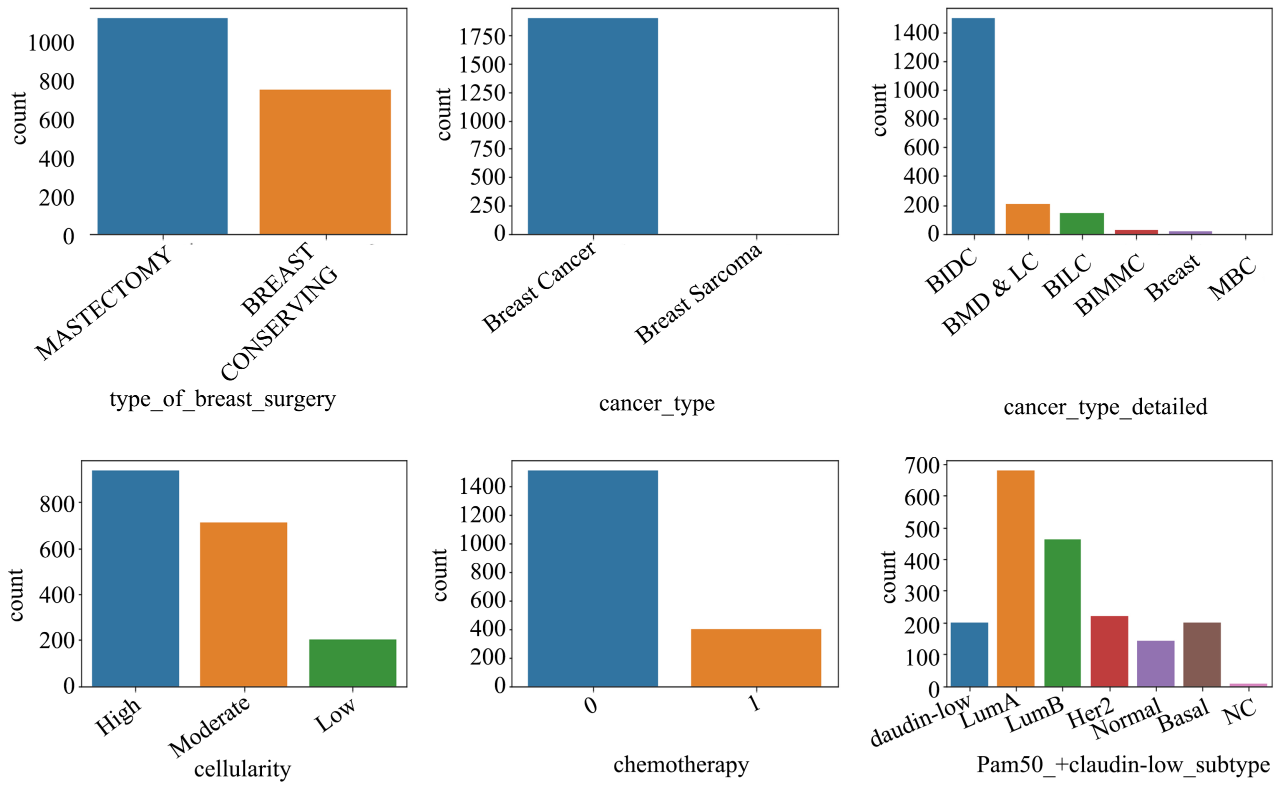

Figure 3. Countplot of categorical clinical features. BIDC, Breast Invasive Ductal Carcinoma; BMD & LC, Breast Mixed Ductal and Lobular Carcinoma; BILC, Breast Invasive Lobular Carcinoma; BIMMC, Breast Invasive Mixed Mucinous Carcinoma; MBC, Metaplastic Breast Cancer.

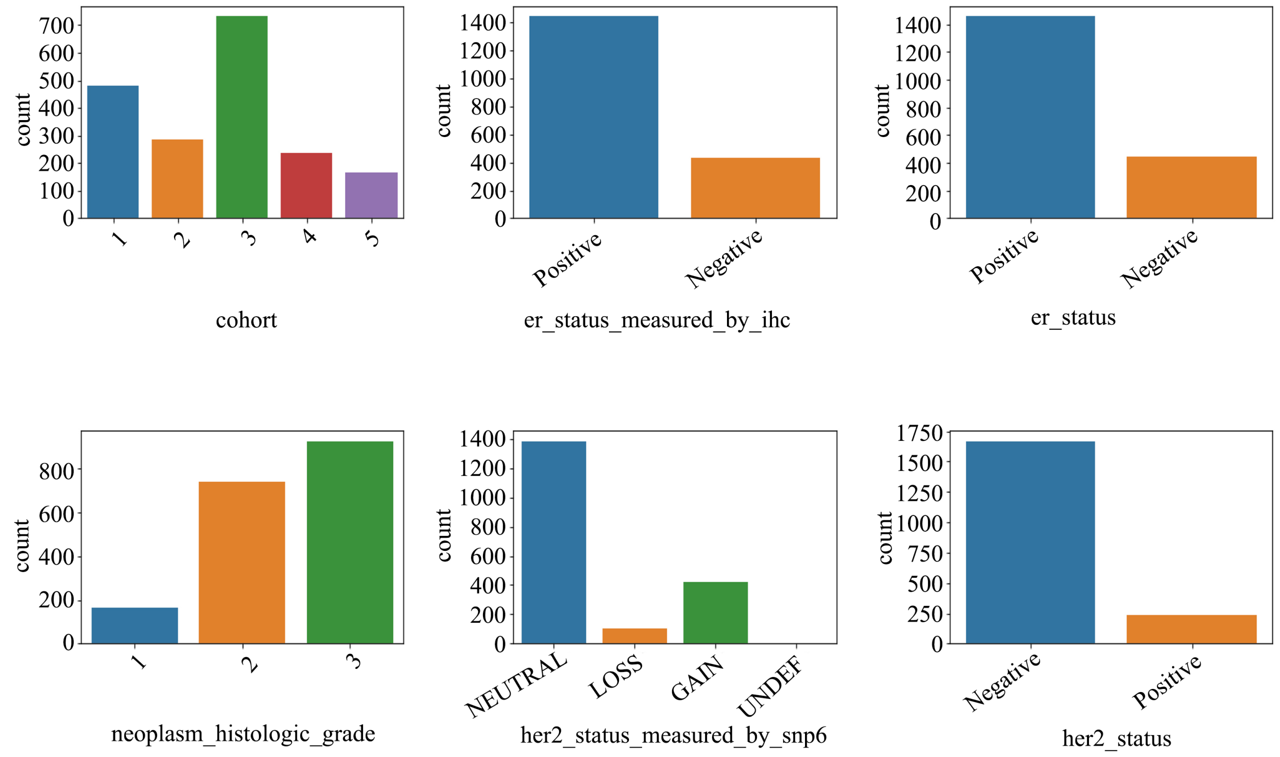

Figure 4. Countplot of categorical clinical features.

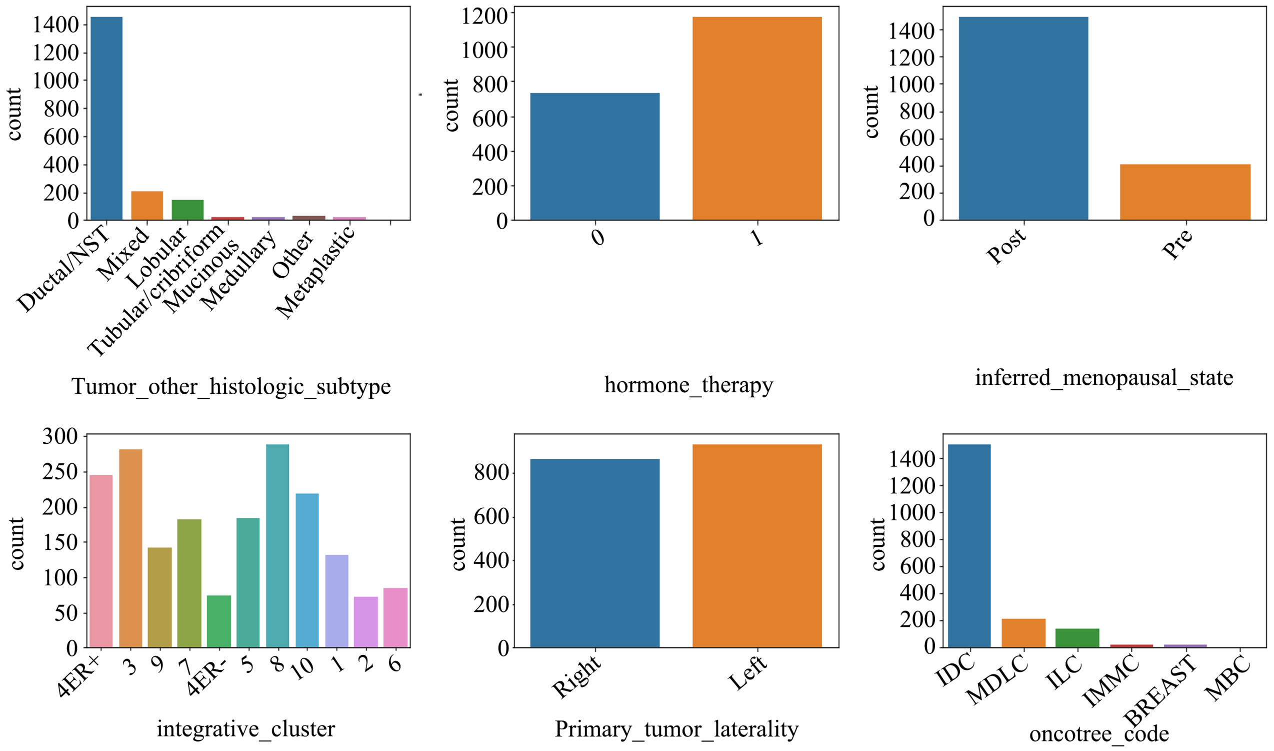

Figure 5. Countplot of categorical clinical features.

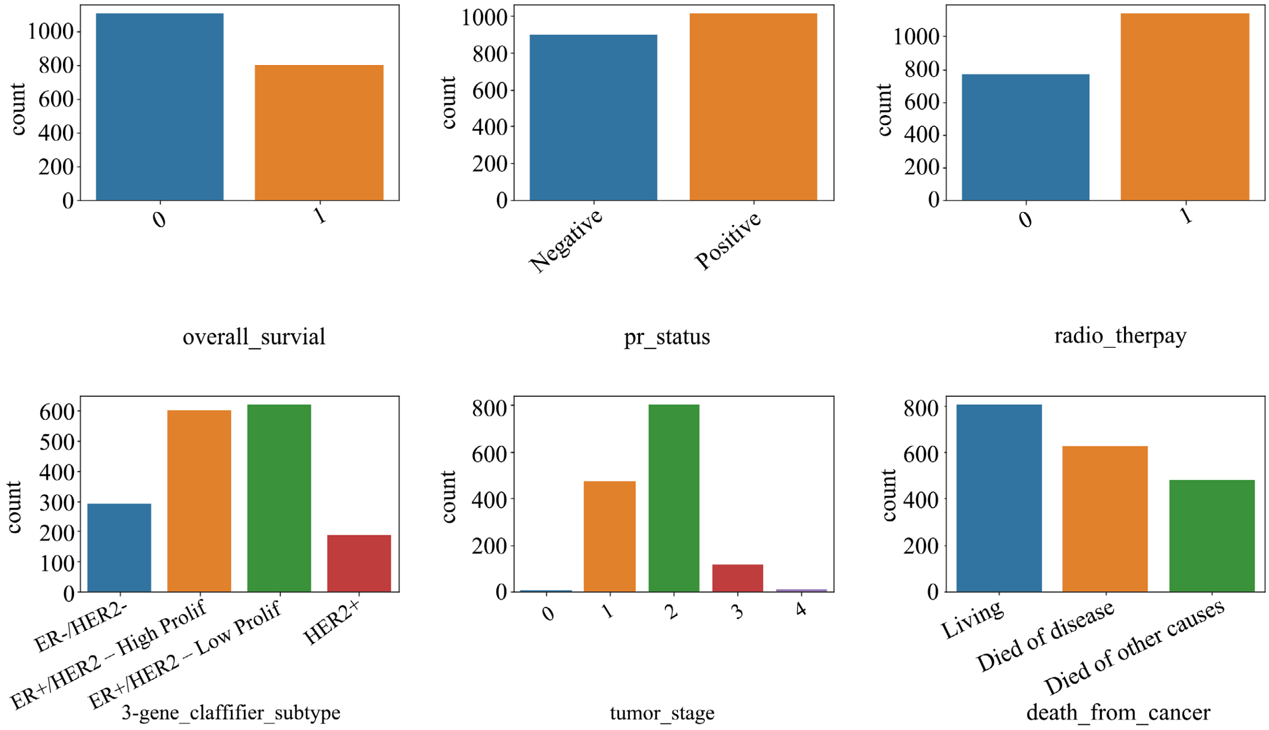

Figure 6. Countplot of categorical clinical features.

Besides clinical features, genetic features and mutation features were also included in analysis (not shown here because of the size).

3.3. Analysis on Feature Based on Target Feature

According to the target feature overall survival, other features were separated into two groups: survival group (overall survival = 1) and non-survival group (overall survival = 0).

For numerical clinical features, they were separated into two groups. Aiming at seeing the difference between the two groups, the boxplot was drawn to show the distribution (Figure 7).

To figure out whether the features of two groups were significantly different or not, Student's t-test were made on numerical features.

For categorical clinical features, the countplots were drawn (Figures 8-11) to see differences between two groups.

χ^2 analysis was used on categorical features to see whether the features of two groups were significantly different or not.

Every genetic feature was separated into two groups using the same criteria, and similar to the way numerical clinical features were tested, Student's t-tests were used on genetic features to see whether the features of two groups were significantly different or not. In all 489 genetic features, 256 of them were significantly different between the two groups, 233 of them were not significantly different between the two groups.

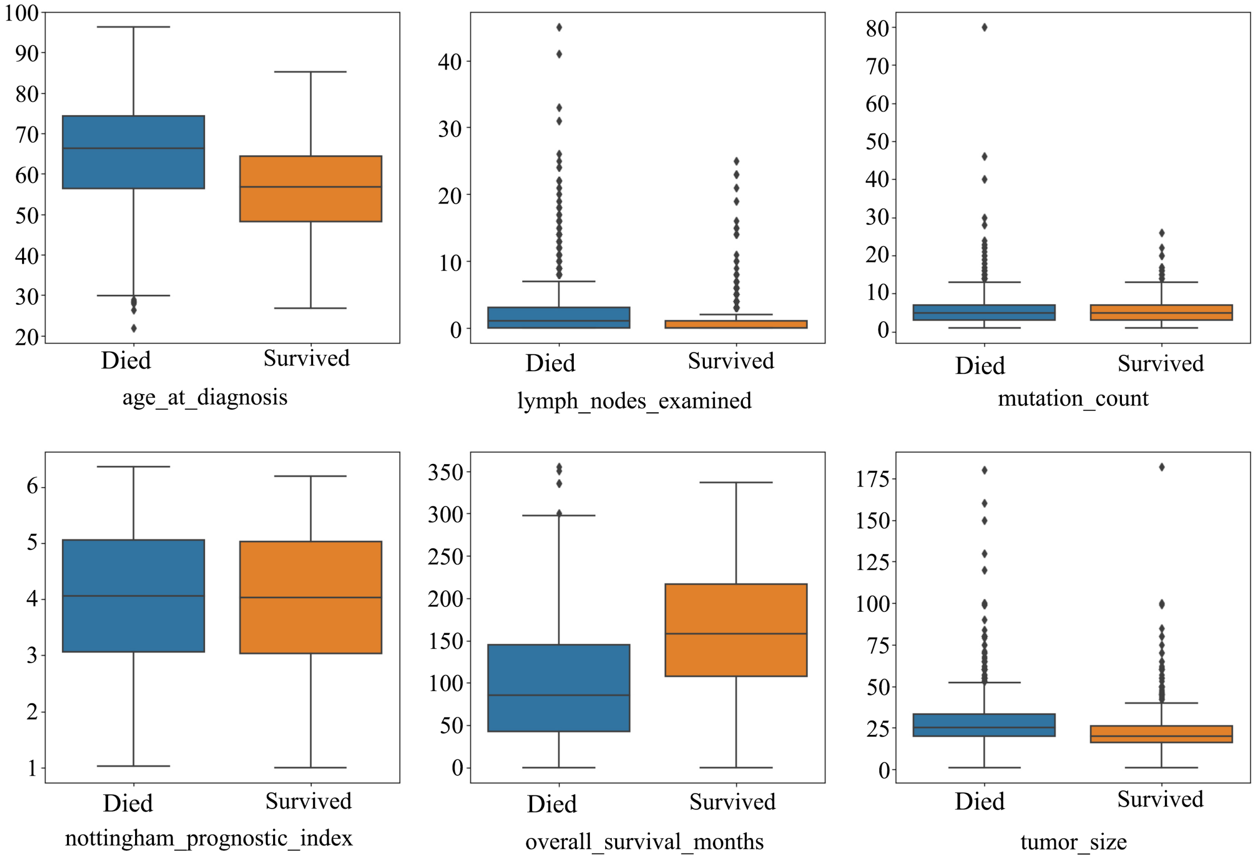

Figure 7. Boxplot of numerical clinical features separated by the target feature.

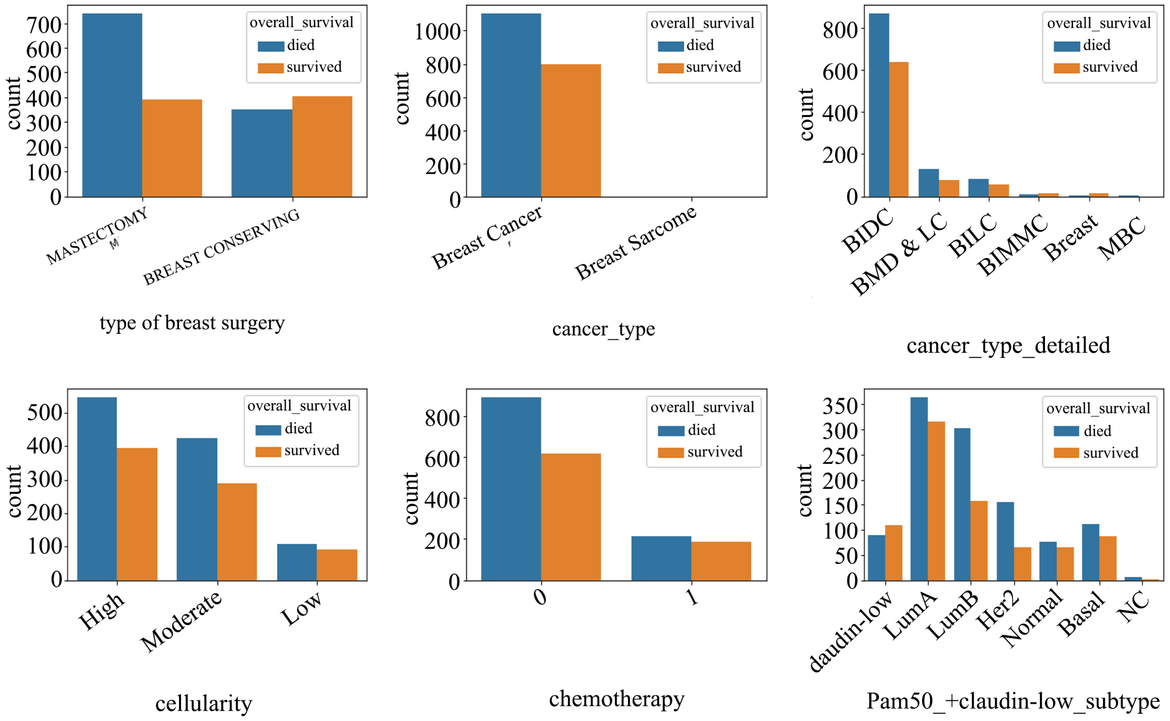

Figure 8. Countplot of categorical clinical features separated by the target feature.

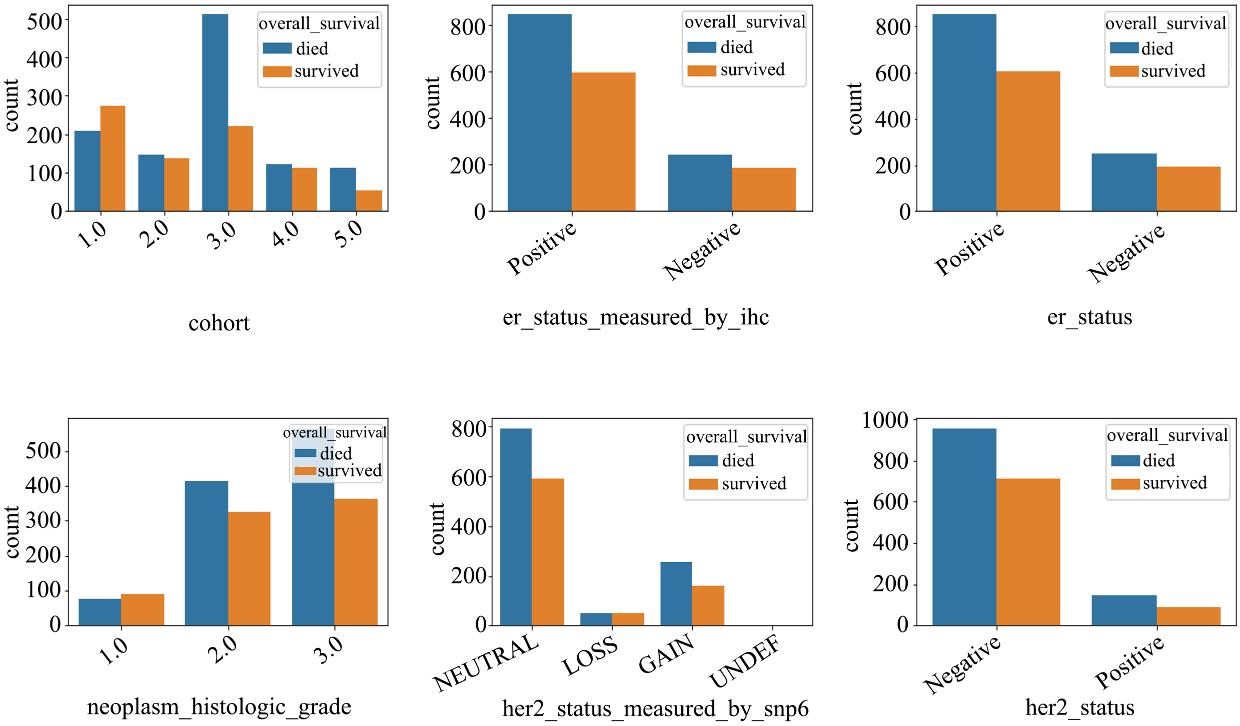

Figure 9. Countplot of clinical categorical features based on the target feature.

For every mutation feature, the countplot was drawn (Figure 12) based on the survival and non-survival group to see differences between them (Not all the mutation features were shown).

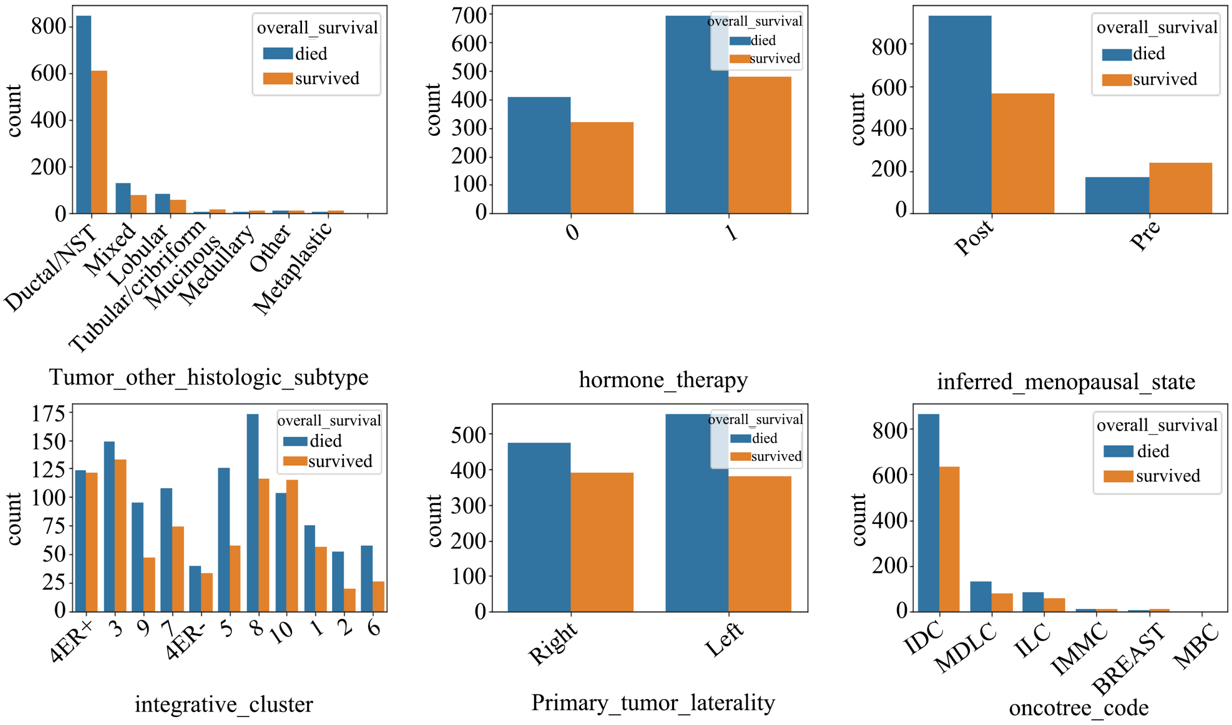

Figure 10. Countplot of categorical clinical features separated by target feature.

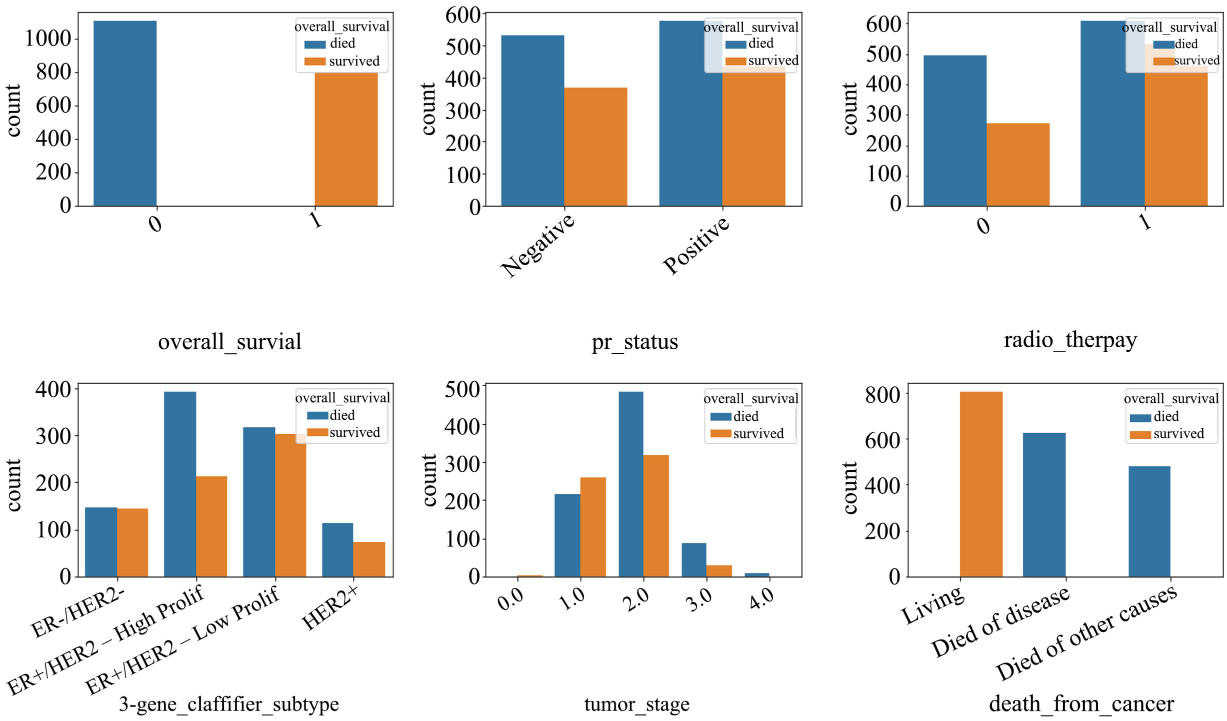

Figure 11. Countplot of categorical clinical features separated by the target feature.

Through the χ^2 analysis, in all 173 mutation features, 15 of them were significantly different between the two groups, 158 of them were not significantly different between them.

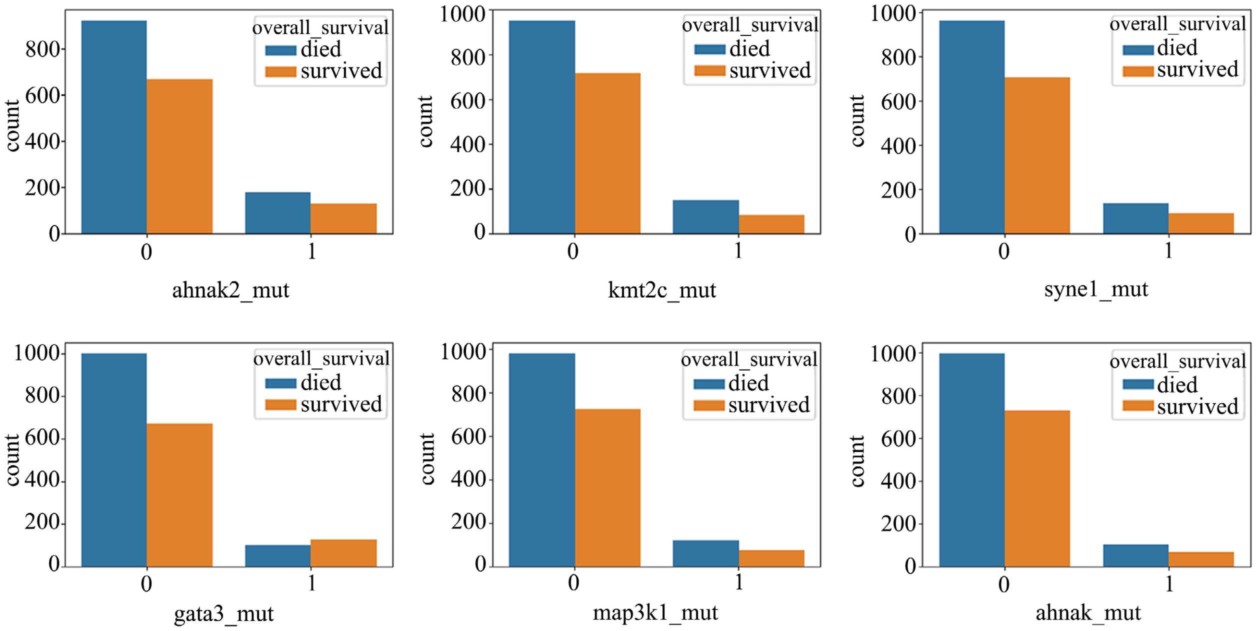

Figure 12. Countplot of mutation features separated by the target feature.

3.4. Model Building

3.4.1. Data Split

In the training part, overall survival was the target feature; all other features were training features. For the 1980 primary breast cancer samples, 20% of them were chosen as test data. What is more, aiming at splitting the data in the same way every time to avoid the bias brought by different split ways, a seeded random state was set.

3.4.2. Model Comparing

In this part, pycaret module was imported to compare different models by using accuracy, AUC, recall and other evaluation indexes. The result was shown in Table 4. Drawing these evaluation indexes of different models (Figure 13), lightgbm (Light Gradient Boosting Machine) model was selected as the training model according to the best comprehensive efficiency. Lightgbm model was used in the subsequent part of training and analysis.

3.4.3. Model Evaluation and Hyperparameter Tuning

After evaluating the model performance, the lightgbm model was used to train the data and the optimal values for the models were determined by hyperparameter tuning.

In the hyperparameter tuning part, learning rate, feature fraction, num leaves, max depth were tuned for better efficiency. Learning rate: the default setting is 0.1, and the general setting is between 0.05 and 0.1. Choosing a smaller learning rate can obtain more stable and better model performance. Feature fraction: the default setting is 1.0, the value of feature_fraction lies between 0.0 and 1.0, if feature fraction is less than 1.0, the model lightgbm will randomly select some features in each iteration. For example, if it is set to 0.8, 80% features will be

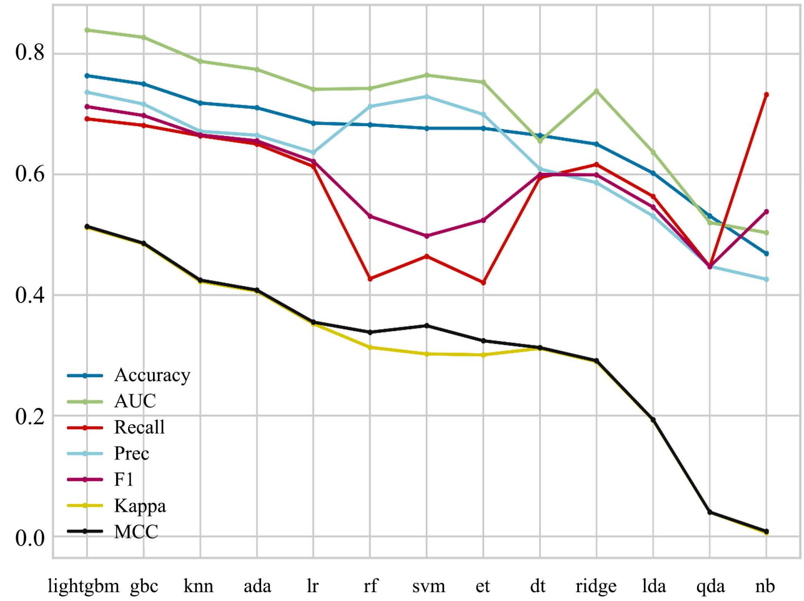

Figure 13. Model performance comparison. lightgbm: Light Gradient Boosting Machine; gbc: Gradient Boosting Classifier; knn: K Neighbors Classifier; ada: Ada Boost Classifier; lr: Logistic Regression; rf: Random Forest Classifier; svm: SVM—Linear Kernel; et: Extra Trees Classifier; dt: Decision Tree Classifier; ridge: Ridge Classifier; lda: Linear Discriminant Analysis; qda: Quadratic Discriminant Analysis; nb: Naive Bayes.

Table 4. Model comparing.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
lightgbm	0.7636	0.8392	0.692	0.7362	0.7123	0.5121	0.5138
gbc	0.7498	0.8271	0.6812	0.7164	0.6975	0.4846	0.4859
knn	0.7183	0.7874	0.664	0.6714	0.6654	0.4226	0.4248
ada	0.7104	0.7739	0.6504	0.6649	0.6556	0.4064	0.4083
lr	0.6849	0.7410	0.6131	0.6365	0.6217	0.3524	0.3551
rf	0.6822	0.7425	0.4271	0.7126	0.5306	0.3133	0.3384
svm	0.6764	0.0000	0.4642	0.7290	0.4980	0.3024	0.3492
et	0.6763	0.7528	0.4209	0.6994	0.5241	0.3009	0.3242
dt	0.6645	0.6553	0.5945	0.6089	0.5999	0.3116	0.3129
ridge	0.6501	0.0000	0.6163	0.5862	0.5992	0.2896	0.2912
lda	0.6021	0.6368	0.5635	0.5309	0.5457	0.1926	0.1934
qda	0.5312	0.5201	0.4473	0.4476	0.4471	0.0403	0.0403
nb	0.4688	0.5034	0.7321	0.4262	0.5386	0.0061	0.0084

LightGBM: Light Gradient Boosting Machine; GBC: Gradient Boosting Classifier; KNN: K Neighbors Classifier; ADA: Ada Boost Classifier; LR: Logistic Regression; RF: Random Forest Classifier; SVM: SVM—Linear Kernel; ET: Extra Trees Classifier; DT: Decision Tree Classifier; Ridge: Ridge Classifier; LDA: Linear Discriminant Analysis; QDA: Quadratic Discriminant Analysis; NB: Naive Bayes.

selected before each tree is trained. By this way, the training can be accelerated. It can also be used to prevent over-fitting. Num leaves: The number of leaf nodes on a tree. The default setting is 31, which cooperates with max depth to empty the shape of the value tree. It is a key parameter that needs to be adjusted, which has a great influence on the model performance. Max depth: Maximum depth of the tree model, which is the most important parameter to prevent over-fitting. It is also the core parameter that needs to be adjusted, which plays a decisive role in model performance and generalization ability.

In the process of tune-up, several values of every model hyperparameter were selected and tuned as following: learning rate = [0.1, 0.3, 0.6], feature fraction = [0.5, 0.8, 1], num leaves = [16, 32, 64], max depth = [−1, 2, 3, 4]. Gridsearchcv was mainly applied in this part. Grid search was to traverse every intersection in the grid to find the best combination. The dimension of the grid was the number of hyperparameters. And by using cross validation = 5, all datasets were divided into five parts, one part was taken as the test set each time without repetition, and the other four parts were trained to create the model. The best model with the least mean-square error was selected out.

According to the part analysis on feature separated by target feature, only some genetic feature and mutation feature were significantly different between the survival and non-survival group. Therefore, for genetic and mutation feature of the new training data, only genetic and mutation features which were significantly different between the two groups were included. By comparing the model performance evaluation parameters of the original and new data, the better one was chosen and used for training.

In summary, for original data without selection and new data with selection, two models (before tune-up and after tune-up) were built separately. In total, four models and their model evaluation index were shown in Table 5.

From the four models selected, the lightgbm model after tune-up which used data with feature selection had the best comprehensive efficiency. The accuracy of the test data set of it was 0.8215, precision of the model was 0.8219, recall was 0.8215, F1 score was 0.8217, Roc_auc score was 0.8708 and the roc_auc curve was shown (Figure 14).

Table 5. Model tune-up and feature selection.

	Data without feature selection		Data with feature selection	
	Before tune-up	After tune-up	Before tune-up	After tune-up
Accuracy of the test data set	0.8084	0.8110	0.8136	0.8215
Precision	0.8086	0.8114	0.8154	0.8219
Recall	0.8084	0.8110	0.8136	0.8215
F1_score	0.8085	0.8112	0.8142	0.8217
Roc_auc score	0.8659	0.8612	0.8620	0.8708

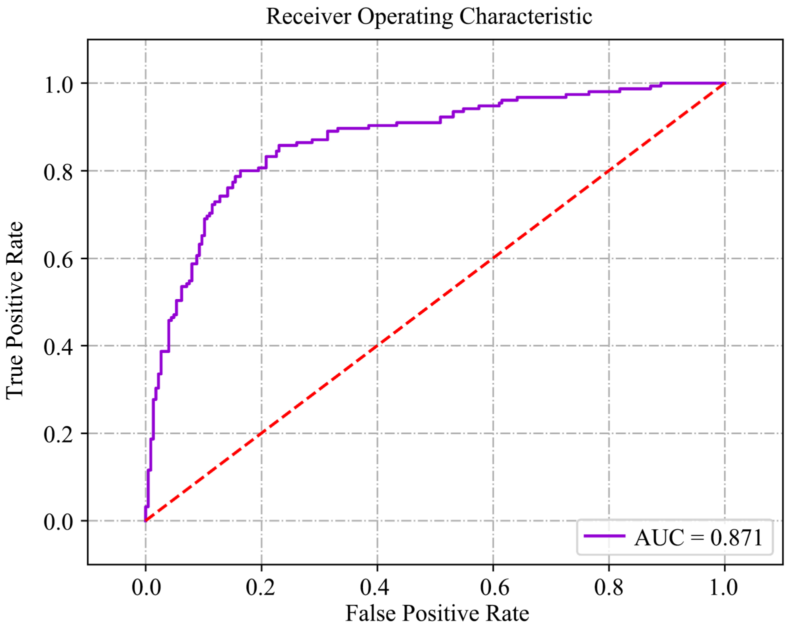

Figure 14. AUC-ROC curve of the lightgbm model.

3.5. Model Interpretation

3.5.1. Feature Importance

The average value of how much contribution each feature had made to each tree in the random forest was calculated and compared (**Figure 15**). In this part, gini index was used as the calculation method of contribution. $Gini(p)=\sum_{k=1}^{K} p_k(1-p_k)$ (K categories, p_k is the sample weight of category k). In this figure, the 5 most important features were age at diagnosis, cohort, Nottingham Prognostic Index, rheb and nr3c1.

3.5.2. Partial Dependence Plot

Partial dependence curve is a visual analysis method to evaluate the influence of a certain dimension feature on the model output, and this method is also independent of the global model. In general, this method gives the average model output score for all possible values of the features input by the model, and a curve is drawn. This method is based on a fundamental assumption: the input of the model is independent.

The trend of partial dependence curve can only represent whether the influence of this feature on the output of the model is positive or negative from the average point of view. Features that can score a larger peak gap tend to be more important.

Through analysis, age at diagnosis, Nottingham Prognostic Index, cohort, rheb and nr3c1 were important to the target feature (**Figure 16**).

4. Discussion

In this article, we hoped to make a basic prediction of the survival of breast cancer patients through the existing data. Data includes clinical data, gene expression

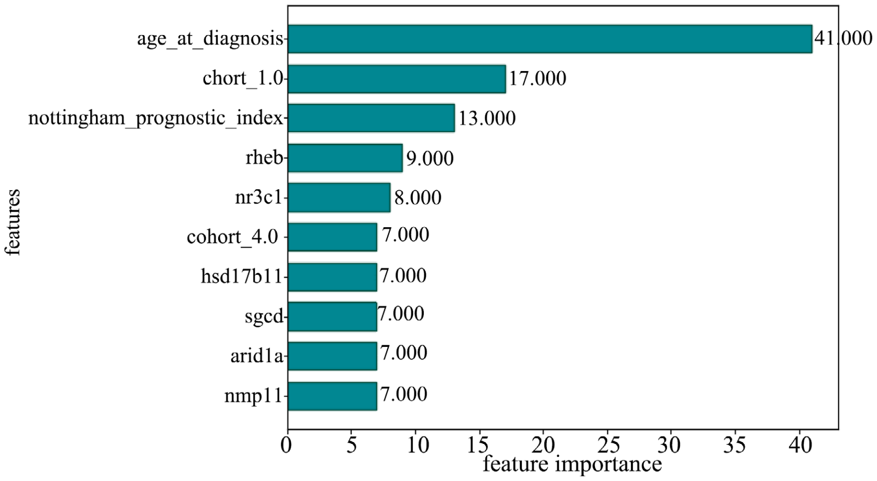

Figure 15. Feature importance plot of the lightgbm model.

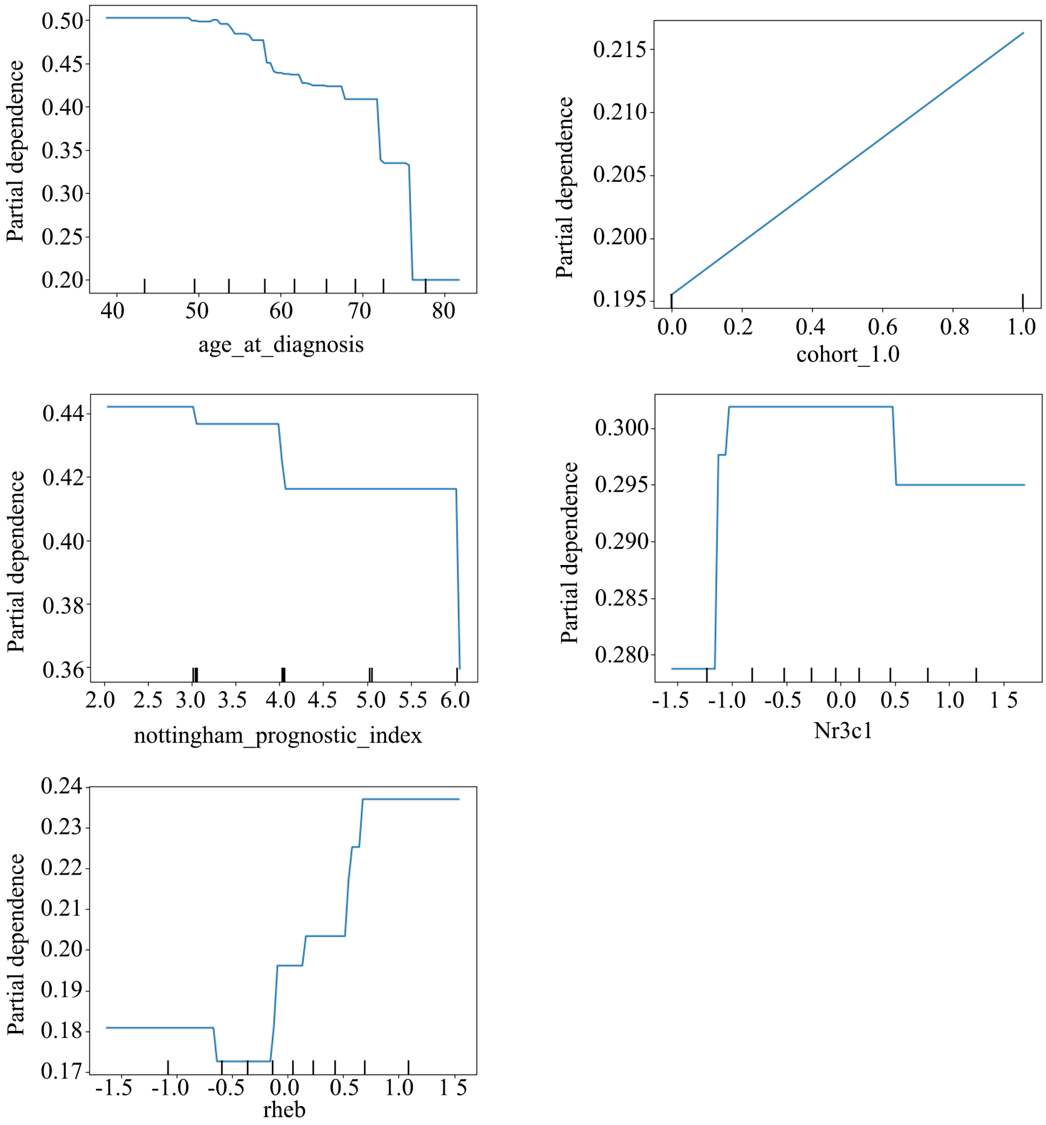

Figure 16. Partial dependence plots of the lightgbm model.

and mutation data. In the process of designing this project, we decided to use machine learning methods to make predictions, hoping to screen out the variables that have great influence on the results under the condition of ensuring the correct classification rate, and provide help for the prevention and treatment of subsequent diseases.

For the common machine learning methods for disease prediction, we mainly optimized them in two aspects: feature selection and method selection. For feature selection, we divided the samples in the data into two groups according to the target features in the previous data analysis and found out the variables with significant differences between the two groups through analysis. In the subsequent training, we tried to train only the variables with significant differences, and through comprehensive analysis of various model evaluation indicators, we found that the prediction performance of the model was improved. In the selection of methods, we imported the pycaret package in python, and use various model evaluation indicators to analyze the common machine learning methods (such as Logistic regression, K neighbors classifier, Naive Bayes, support vector machine (SVM), Decision Tree Classifier) and ensemble learning methods (such as Random Forest, Light Gradient Boosting Machine, Gradient Boosting Classifier) were analyzed and compared, and the model with the best performance was selected and used for final prediction [10] [11] [12]. Among common methods, although depending on different datasets and parameter selection each method performs in a different way, K neighbors classifier gives the best results and support vector machine is the most suitable for predicting [19]. However, while the irrelevant features increase, the computation time of the support vector machine increases, and the error rate is also higher [20]. Among ensemble learning methods, many methods have been proved to have great effect on the aspect of predicting breast cancer [21] [22] [23]. Therefore, these models were taken into our consideration, compared and selected. According to the survival situation of patients in training data, our method finally divides breast cancer patients into the safe group (survival group) and the risky group (non-survival group). We hope that this classification can be used as a reference to help physicians diagnose the severity of breast cancer. At the same time, we also hope that this classification can enable patients to be treated in groups, thus making the treatment plan more targeted and efficient, improving the cure rate of diseases.

In addition, we analyzed the feature importance of the final results, and selected five features that have the greatest impact on the final prediction results of the model statistically, namely, age at diagnosis, Nottingham Prognostic Index, cohort, rheb and nr3c1. The first three important features are common and easy to obtain in the clinic, so we can pay more attention to them in the future disease diagnosis and treatment. The other two are related to gene expression. In the present study, these genes have been proven to have impacts on cancer. The gene rheb has a strong relationship with BRAF; this relationship is affected by the Y35N point mutation, which can cause cellular cancer transformation [24]. The gene nr3c1 encodes glucocorticoid receptor (GR), and GR binds to cortisol, which has an elevated level in breast cancer patients [25]. However, because they are not common therapeutic genes for breast cancer, we can continue to explore whether they have an impact on the formation of breast cancer. This study shows that automated group therapy might be helpful to physicians clinically. And more accurate and efficient classification indexes of diseases can be explored in follow-up study. These indexes will play an important role in improving the efficiency of diagnosis and prognosis.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] WHO (2023) Global Breast Cancer Initiative Implementation Framework: Assessing, Strengthening and Scaling up of Services for the Early Detection and Management of Breast Cancer: Executive Summary.

[2] Luo, J., *et al.* (2022) Etiology of Breast Cancer: A Perspective from Epidemiologic Studies. *Journal of the National Cancer Center*, **2**, 195-197. https://doi.org/10.1016/j.jncc.2022.08.004

[3] Pfob, A., *et al.* (2021) Identification of Breast Cancer Patients with Pathologic Complete Response in the Breast after Neoadjuvant Systemic Treatment by an Intelligent Vacuum-Assisted Biopsy. *European Journal of Cancer*, **143**, 134-146. https://doi.org/10.1016/j.ejca.2020.11.006

[4] Lenkinski, R.E. (2022) Improving the Accuracy of Screening Dense Breasted Women for Breast Cancer by Combining Clinically Based Risk Assessment Models with Ultrasound Imaging. *Academic Radiology*, **29**, S8-S9. https://doi.org/10.1016/j.acra.2021.09.019

[5] Jalalian, A., *et al.* (2013) Computer-Aided Detection/Diagnosis of Breast Cancer in Mammography and Ultrasound: A Review. *Clinical Imaging*, **37**, 420-426. https://doi.org/10.1016/j.clinimag.2012.09.024

[6] Ruiz, A., *et al.* (2018) Surgical Resection versus Systemic Therapy for Breast Cancer Liver Metastases: Results of a European Case Matched Comparison. *European Journal of Cancer*, **95**, 1-10. https://doi.org/10.1016/j.ejca.2018.02.024

[7] Ward, K.A., *et al.* (2023) Long-Term Adherence to Adjuvant Endocrine Therapy Following Various Radiotherapy Modalities in Early Stage Hormone Receptor Positive Breast Cancer. *Clinical Breast Cancer*, **23**, 369-377. https://doi.org/10.1016/j.clbc.2023.01.012

[8] Jacobs, A.T., *et al.* (2022) Targeted Therapy for Breast Cancer: An Overview of Drug Classes and Outcomes. *Biochemical Pharmacology*, **204**, Article ID: 115209. https://doi.org/10.1016/j.bcp.2022.115209

[9] Nemade, V. and Fegade, V. (2023) Machine Learning Techniques for Breast Cancer Prediction. *Procedia Computer Science*, **218**, 1314-1320. https://doi.org/10.1016/j.procs.2023.01.110

[10] Khandezamin, Z., Naderan, M. and Rashti, M.J. (2020) Detection and Classification of Breast Cancer Using Logistic Regression Feature Selection and GMDH Classifier. *Journal of Biomedical Informatics*, **111**, Article ID: 103591. https://doi.org/10.1016/j.jbi.2020.103591

[11] Pratheep, K.P., *et al.* (2021) An Efficient Classification Framework for Breast Cancer Using Hyper Parameter Tuned Random Decision Forest Classifier and Bayesian Optimization. *Biomedical Signal Processing and Control*, **68**, Article ID: 102682. https://doi.org/10.1016/j.bspc.2021.102682

[12] Chakravarthy, S.S.R., Bharanidharan, N. and Rajaguru, H. (2023) Deep Learning-Based Metaheuristic Weighted K-Nearest Neighbor Algorithm for the Severity Classification of Breast Cancer. *IRBM*, **44**, Article ID: 100749. https://doi.org/10.1016/j.irbm.2022.100749

[13] Pereira, B., *et al.* (2016) The Somatic Mutation Profiles of 2,433 Breast Cancers Refines Their Genomic and Transcriptomic Landscapes. *Nature Communications*, **7**, Article No. 11479. https://doi.org/10.1038/ncomms11908

[14] Sundquist, M., *et al.* (1999) Applying the Nottingham Prognostic Index to a Swedish breast cancer population. *Breast Cancer Research and Treatment*, **53**, 1-8. https://doi.org/10.1023/A:1006052115874

[15] Jones, C. and Lancaster, R. (2018) Evolution of Operative Technique for Mastectomy. *Surgical Clinics of North America*, **98**, 835-844. https://doi.org/10.1016/j.suc.2018.04.003

[16] Guiu, S., *et al.* (2012) Molecular Subclasses of Breast Cancer: How Do We Define Them? The IMPAKT 2012 Working Group Statement†. *Annals of Oncology*, **23**, 2997-3006. https://doi.org/10.1093/annonc/mds586

[17] Mazouni, C., *et al.* (2010) Is Quantitative Oestrogen Receptor Expression Useful in the Evaluation of the Clinical Prognosis? Analysis of a Homogeneous Series of 797 Patients with Prospective Determination of the ER Status Using Simultaneous EIA and IHC. *European Journal of Cancer*, **46**, 2716-2725. https://doi.org/10.1016/j.ejca.2010.05.021

[18] Liu, D. and Zhou, K. (2020) BRAF/MEK Pathway Is Associated with Breast Cancer in ER-Dependent Mode and Improves ER Status-Based Cancer Recurrence Prediction. *Clinical Breast Cancer*, **20**, 41-50, e8. https://doi.org/10.1016/j.clbc.2019.08.005

[19] Rana, M., *et al.* (2015) Breast Cancer Diagnosis and Recurrence Prediction Using Machine Learning Techniques. *International Journal of Research in Engineering and Technology*, **4**, 372-376. https://doi.org/10.15623/ijret.2015.0404066

[20] Memon, M., *et al.* (2019) Breast Cancer Detection in the IOT Health Environment Using Modified Recursive Feature Selection. *International Journal of Wireless and Mobile Computing*, **2019**, Article ID: 5176705. https://doi.org/10.1155/2019/5176705

[21] Minnoor, M. and Baths, V. (2023) Diagnosis of Breast Cancer Using Random Forests. *Procedia Computer Science*, **218**, 429-437. https://doi.org/10.1016/j.procs.2023.01.025

[22] Jiang, Z., *et al.* (2021) A Light Gradient Boosting Machine-Enabled Early Prediction of Cardiotoxicity for Breast Cancer Patients. *International Journal of Radiation Oncology, Biology, Physics*, **111**, e223. https://doi.org/10.1016/j.ijrobp.2021.07.771

[23] Abbasniya, M.R., *et al.* (2022) Classification of Breast Tumors Based on Histopathology Images Using Deep Features and Ensemble of Gradient Boosting Methods. *Computers and Electrical Engineering*, **103**, Article ID: 108382. https://doi.org/10.1016/j.compeleceng.2022.108382

[24] Heard, J.J., *et al.* (2018) An Oncogenic Mutant of RHEB, RHEB Y35N, Exhibits an Altered Interaction with BRAF Resulting in Cancer Transformation. *BMC Cancer*,

18, Article No. 69. https://doi.org/10.1186/s12885-017-3938-5

[25] Gandhi, S., *et al*. (2020) Contribution of Immune Cells to Glucocorticoid Receptor Expression in Breast Cancer. *International Journal of Molecular Sciences*, **2**, Article No. 4635. https://doi.org/10.3390/ijms21134635