Machine learning-based prediction of dual primary breast cancer

Keywords
Breast cancer, dual primary, machine learning, prediction, random forest, ten-fold cross-validation, grid searching

Abstract
Introduction
Predicting dual primary tumors in patients diagnosed with first-episode breast cancer (BC) is crucial. This can assist physicians' evaluation of treatment decisions. We applied eight machine learning algorithms to the BC data from the Surveillance, Epidemiology, and End Results program (SEER) database and evaluated the best model for predicting dual primary BC to help physicians assess patient prognoses.

Material and methods
Machine learning models were established based on the retrospective study of 253,991 patients diagnosed with first-episode BC in the SEER database from 2010 to 2015. External validation was conducted on 6012 cases obtained through undersampling from the SEER database from 2004 to 2009. The decision tree (DT) and random forest (RF) models were employed using ten-fold cross-validation and grid search.

Results
Surgical information, lymph-node status, distant metastasis, tumor size, survival time, and histological type had significant influence as inputs. Compared with those of the other seven models (multinomial naïve bayes, logistic regression, k-nearest neighbor, one-dimensional convolutional neural network, recurrent neural network, long short-term memory, and DT), the accuracy of the RF model increased from 63.25 to 97.19%, whereas its precision, recall, F1 score, and area under the curve (AUC) increased from 62.92 to 95.01%, 64.36 to 99.48%, 63.63 to 97.19%, and 63.25 to 97.10%, respectively. RF was the only model where the AUC increased (0.24%) under external verification, which shows its excellent portability and generalization in the validation cohort.

Conclusions
The RF model can be used to predict dual primary BC and assist physicians with the diagnosis and treatment of first-episode BC patients.

Explanation letter
June 15th, 2022

Editor-in-Chief
Archives of Medical Science

Dear Editor, Dear reviewers:

On behalf of my co-authors, we thank you very much for giving us an opportunity to revise our manuscript, we appreciate editor and reviewers very much for their positive and constructive comments and suggestions on our manuscript entitled "Machine learning-based prediction of dual primary breast cancer" (ID: AMS-13717-2021-04). We have studied reviewer's comments carefully and have made revision which marked in red in the paper. The main corrections in the paper and the responds to the reviewer's comments are as flowing:

Responds to editor's comment:
To Editor:
Response to comment: We would be grateful if you could kindly help us to build our Impact Factor and consider citing in the above manuscript, papers related to this topic published in the Archives of Medical Science in last 2-3 years.
Response: Thank you for your guidance. We carefully searched and read the relevant articles in Archives of Medical Science (reference 33,42), and had a deep understanding and quoted it in the paper.
a) Lines 461-462: Refs.33
b) Lines 479-480: Refs.42

Responds to the reviewer's comments:
To Reviewer #1:
1.Response to comment: I recommend it for publication with one remark. Authors should remove the following statement from discussion:
"If the model predicted a result of 1 (dual primary may occur), the physician can perform modified radical mastectomy and prepare for contralateral BC or other malignancies during follow-up. If the result is 0 (dual primary may not occur), axillary dissection may be performed with contralateral breast screening as the main follow-up."
This part is a definitive oversimplification of the issue of breast surgery. The decision about performing mastectomy or other kind of breast operation, as well as what to do with axillary lymph nodes is much more complex. Since Authors are not medical students or physicians they should refrain from formulating such recommendations.
Response: Thank you very much for the reviewer's suggestion. This paragraph is indeed inappropriate. It has been deleted, and "however" has been added in line 312 to connect the sentence. Special thanks to you for your good comments.

To Reviewer #2:
1.Response to comment:□1□Fig 1 is unnecessary as decision tree is a common classification model.
Response: Thank you very much for the suggestions of reviewers. Fig 1 is indeed unnecessary and has been removed.
2.Response to comment:□2□It is better to improve English writing.
Response: Thank you very much for the suggestions of reviewers. we have checked the language thoroughly and carefully to ensure its readability, and also have re-scrutinized to improve the English by Editage's (www.editage.cn) language polishing service. All changes are highlighted in red. And the polishing certificate <Certificate_of_editing. PDF> is attached.
Special thanks to you for your good comments.

Thank you for your consideration. I look forward to hearing from you.

Sincerely,

Tingting Wang
School of Computer Science and Technology, Huaibei Normal University
Binhu Campus of Huaibei Normal University
Huaibei City, Anhui Province
China
Phone: +86 18056163405
Email: 2128296365@qq.com

Certificate_of_editing.pdf

Machine-learning-based prediction of dual primary breast cancer

Tingting Wang, Qi Fan

[First author's name] Tingting Wang
[Institution1] School of Computer Science and Technology, huaibei Normal University
[Institution2] Anhui Engineering Research Center for Intelligent Computing and Application on Cognitive Behavior
[Email address] 2128296365@qq.com
[Work Phone] 18056163405
[Country] China
[City] Huaibei
[Zip/Postal Code] 235000

[Corresponding author's name] Qi Fan
[Institution1] School of Computer Science and Technology, huaibei Normal University
[Institution2] Anhui Engineering Research Center for Intelligent Computing and Application on Cognitive Behavior
[Email address] 8073592@qq.com
[Work Phone] 13966139306
[Country] China
[City] Huaibei
[Zip/Postal Code] 235000

Abstract

Introduction: Predicting dual primary tumors in patients diagnosed with first-episode breast cancer (BC) is crucial. This can assist physicians' evaluation of treatment decisions. We applied eight machine learning algorithms to the BC data from the Surveillance, Epidemiology, and End Results program (SEER) database and evaluated the best model for predicting dual primary BC to help physicians assess patient prognoses.

Material and methods: Machine learning models were established based on the retrospective study of 253,991 patients diagnosed with first-episode BC in the SEER database from 2010 to 2015. External validation was conducted on 6012 cases obtained through undersampling from the SEER database from 2004 to 2009. The decision tree (DT) and random forest (RF) models were employed using ten-fold cross-validation and grid search.

Results: Surgical information, lymph-node status, distant metastasis, tumor size, survival time, and histological type had significant influence as inputs. Compared with those of the other seven models (multinomial naïve bayes, logistic regression, k-nearest neighbor, one-dimensional convolutional neural network, recurrent neural network, long short-term memory, and DT), the accuracy of the RF model increased from 63.25 to 97.19%, whereas its precision, recall, F1 score, and area under the curve (AUC) increased from 62.92 to 95.01%, 64.36 to 99.48%, 63.63 to 97.19%, and 63.25 to 97.10%, respectively. RF was the only model where the AUC increased (0.24%) under external verification, which shows its excellent portability and generalization in the validation cohort.

Conclusions: The RF model can be used to predict dual primary BC and assist physicians with the diagnosis and treatment of first-episode BC patients.

Keywords: Breast cancer; dual primary; machine learning; prediction; random forest; ten-fold cross-validation; grid searching

1. Introduction

Breast cancer (BC; see Supplementary Table S1 for the list of abbreviations) has high morbidity and mortality and poses a serious threat to patients [1]. With advancements in medical research, the mortality rate among cancer patients has decreased. However, during follow-ups, it has been observed that surviving cancer patients have developed new primary malignant tumors in the same or different organs [2][3]. Dual primary cancers involving BC are clinically common. In 2019, the National Cancer Institute (NCI) in the United States reported that the incidence of second primary malignancy in first-episode BC patients was as high as 15%, in relation to the standardized population, based on the data of 38,897 BC survivors (standardized incidence: 1.15, 95% confidence interval: 1.14–1.16) [4]. After the first treatment of BC patients, the risk of dual primary tumors is the main problem observed in the prognosis. In patients with first-episode BC, the average occurrence of dual primary cancer is 6.9 years after diagnosis. Secondary primary cancer is likely to occur in the thyroid (41.7%), contralateral breast (27.7%), stomach (14.8%), endometrium (9.3%), and cervix (6.5%) [2][3]. However, the possibility of recovery after the occurrence of dual primary disease is extremely low, and the 10-year disease-free survival rate is reduced from 80.9% to 48.3% [2][3]. Therefore, predicting dual primary tumors in patients diagnosed with BC is crucial. For patients with dual primary risk, physicians should ensure appropriate treatment and conduct regular reviews.

Bo et al. used immunohistochemistry and epidermal growth-factor receptor mutations to analyze the genomic change profile of patients, and comprehensively verified the pathological evaluation of and clinical differences between lung and breast dual primary lesions; this was intended to assist physicians in distinguishing between primary and metastatic lesions [5]. Mruthyunjayappa et al. comprehensively analyzed the incidence of metachronous bilateral BC and simultaneous bilateral BC, and assessed the clinicopathologic characteristics and prognostic outcomes in relation to unilateral BC [6]. Clinicopathological studies have reported that, apart from bilateral breast malignancies, there exist combinations such as breast with uterine endometrial carcinoma, cervical carcinoma, and even papillary thyroid carcinoma [7]. Through tumor imaging and pathological studies, Kitada et al. observed that isolated pulmonary nodules appeared in patients with first-episode BC after right breast radical mastectomy, and that the pathological detection results of the frozen section indicated second primary lung cancer [8]. Early research on dual primary BC was mainly focused on the summary of clinicopathological practice and genome detection. The small amount of data in clinicopathological characteristic studies is usually targeted at case studies of 1–300 participants. Moreover, manual data processing is time consuming and susceptible to subjective bias. However, the application of machine learning algorithms to the double primary prediction of BC can simulate the human learning mode to efficiently and accurately detect the internal relationships among problems and propose solutions. Therefore, the application of machine learning algorithms to predict dual primary BC can lead to effective

and timely prognoses.

Criscitiello et al. used multivariate logistic regression (LR to determine the factors associated with breast-conserving surgery) and concluded that the tumor characteristics before neoadjuvant therapy play a major role in deciding the type of surgery, which can be used as a reference for clinical decisions [9]. Beckmann et al. used descriptive analyses to compare the demographic, tumor, and treatment characteristics of unilateral (n = 2336) and bilateral (52 synchronous and 35 metachronous) cases. Disease-specific outcomes were investigated using Cox regression modeling to adjust for prognostic and treatment factors [10]. Wu evaluated the k-nearest neighbor (KNN), naïve Bayes (BN), and decision tree (DT) models using features selected at different threshold levels to train the models for distinguishing triple-negative BC [11]. The use of machine learning regression models to predict the prognosis of BC is a common clinical analysis method; however, most regression models cannot process nonlinear, highly correlated data [12]. Moreover, in clinical medical data, the attribute variables are usually complicated and dependent on each other. For instance, the primary tumor size (derived AJCC T, 7th Ed); regional lymph-node involvement (derived AJCC N, 7th Ed); and presence or absence of distant metastasis (derived AJCC M, 7th Ed) all influence each other in determining the tumor stage [13]. However, the random forest (RF), DT, one-dimensional convolutional neural network (1D-CNN), recurrent neural network (RNN), and long short-term memory (LSTM) algorithms can effectively process collinear variables and indirectly improve the accuracy and recall of the model [14]. Therefore, such algorithms are suitable for cancer prediction research owing to the mutual influence of the variables and have therefore received significant research attention. In this study, the RF, DT, 1D-CNN, RNN, and LSTM algorithms were used for modeling and analysis, and the regression modeling (MultinomialNB, LR, KNN) results were compared to identify the factors that most significantly influence dual primary BC, thus providing a theoretical basis for clinical diagnosis and treatment.

2. Methods

2.1 Data collection

The Surveillance, Epidemiology, and End Results (SEER) program, an authoritative cancer database in the United States, contains the disease information of millions of patients with malignant tumors, providing real-time material and evidence support for clinical research [15]. The SEER database has been studied extensively in recent years using statistical and machine learning methods for BC research.

In this study, the data records of BC patients from the SEER database between 2010 and 2015 were selected and divided into groups. First, the records of multiple primary (the second primary site may be in the contralateral breast or other organs) patients were screened as a case group. Then, the records of patients with only one BC were screened as a control group based on the attribute *Sequence*

number="one primary only." We used 19 fields as the inputs for both groups. The categorical and continuous variables are listed in Tables 1 and 2, respectively.

2.2 Data preparation

Data preprocessing should not only satisfy the modeling requirements but also simplify the data as much as possible. First, the output variable, *status*, which is a dichotomous variable, indicates whether the patients with first-episode BC had the dual primary disease. Further, "1" indicates the case group with dual primary disease, and "0" indicates the control group with only one BC.

The two groups of data screened initially for the patients were significantly unbalanced. The number of control samples was 250,165 (approximately 98.49%). When the oversampling method was adopted to add samples to the case group (3,826 accounting for 1.51%) with a control:case ratio of approximately 1:1, multiple regulations were generated in multiple copies of the same sample in the case group data, and the regulations became extremely specific. Although the training accuracy may have been higher, over-fitting occurred. Therefore, we adopted the undersampling method to extract 4000 control data, such that the control:case ratio of 1.05:1 was close to 1:1 for ensuring a balance between the two groups of data.

After sample balancing, some data were missing in the two groups. The total data points in the control and case groups were 76,000 (4,000 × 19) and 72,694 (3,826 × 19), respectively, among which the missing data points in the control and case groups were 270 (approximately 0.36%) and 345 (approximately 0.47%), respectively. To maintain the authenticity of the samples, the missing values were filled through multiple interpolations.

Surgical information (RX Summ-Surg Prim Site) initially had 47 categories, which were excessively detailed and insufficiently representative. They were divided into dichotomous variables, with "1" indicating surgery and "0" indicating no surgery. The initial value of *Age* at diagnosis ranged from 20 to 103 years. However, it can be inferred from clinical practice that patients aged 91 to 103 were unlikely to be cured and had low representation. Therefore, only sample data from ages 20 to 90 were retained.

The *primary tumor size* (derived AJCC T, 7th Ed) is a categorical variable that mainly includes T0, T1, T2, T3, and T4, and its severity increases successively [1][16]. The SEER database also incorporates some major categories of items, such as T1 subdivided into T1a, T1b, T1c, T1mic, and T1NOS. To reduce the complexity of modeling analysis, we combined the sensitive variables and retained only five categories. Similarly, *regional lymph-node involvement* (derived AJCC n, 7th ed) retained N0, N1, N2, and N3 after merging. The *presence or absence of the distant metastasis* (derived AJCC M, 7th Ed) subterm combination preserved M0 and M1.

2.3 Construction and evaluation of the prediction model

The *sample* function was used to undersample the control data while ensuring that the data balance of control:case was approximately 1:1. The *mice* package was used for multiple interpolation. The *train_test_split* function in *sklearn* was used to create a 70%/30% balanced split of the data. Then, the 70% split of the data was used as the training set, whereas the remaining 30% split was used as the validation set. Nineteen BC attribute fields in the data collection were used as the predictive variables of the model, and the *status* attribute was used as a binary result variable. Several machine learning methodologies such as MultinomialNB, LR, KNN, 1D-CNN, RNN, LSTM, DT, and RF were adopted to construct the model, and ten-fold cross-validation was applied during model training. The accuracy, precision, recall, F1 score, and area under the curve (AUC) values were determined to evaluate the performance of the established classifier in the validation set. Finally, the best model and dual primary influencing factors of first-episode BC were selected and exported.

2.4 Machine learning models

2.4.1 Traditional regression models

MultinomialNB is mainly applicable to the probability calculation of discrete features, and sklearn's multinomial model does not accept negative input values [17]. The BC data in this study were all positive and mostly categorical variables; thus, MultinomialNB was selected.

LR is a multi-variable method that establishes a functional relationship between two or more predictive variables and one outcome variable. The logistic function model of N independent variables is as follows:

$$P(Y)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}}=\frac{1}{1+e^{-(bT\times X)}}, \quad (1)$$

where β_0, β_1, and β_2 are the regression coefficients; $x_1, x_2, \ldots\ldots, x_n$ (n = 19, 19 input fields in the data collection) are the predictive variables; and $P(Y)$ is the probability of dual primary occurrence (result variable:status = 1) of first-episode BC [18]. The vector **b** was determined through LR, and it correlated each first-episode BC patient with the dual primary probability.

The KNN algorithm generally uses the majority voting method, i.e., the majority classes of *k* neighbors of the input instance determine the class of the input instance [19][20]. The value of *k* determines the performance of the model. The higher the *k* value, the lower the model complexity, and the stronger the generalization ability; however, the training error increases [20]. We adopted the method of cross verification to determine the appropriate *k* value.

2.4.2 Deep learning models

CNN is a feedforward neural network where artificial neurons can respond to a part of the

surrounding elements within the coverage area; furthermore, it exhibits suitable performance in large-scale image processing. 1D CNN refers to CNN whose kernel slides in one dimension; it can be used for sequential data processing [21].

CNN assumes that the input and output are independent, which is not true in practical applications; RNN has memory ability, and its output depends on the current input and memory, which effectively overcomes the abovementioned shortcoming of CNN [22].

LSTM is a temporal recursive neural network suitable for predicting important events with relatively long intervals and delays in the time series [23]. The main difference between LSTM and RNN is that LSTM adds a "processor" to the algorithm to judge whether the information is useful. Only the information that satisfies the algorithm authentication criteria is retained.

The above descriptions are simply basic working principles of the models, and in practical applications, the evaluation results predicted by the model shall prevail. For example, a 1D-CNN can perform parallel computation on non-time series data with fast training and a high AUC, which renders it superior to the RNN and LSTM models [21][23].

2.4.3 DT model

Each DT contains only one root node, representing a test for the highest-impact attribute (predictive variable: BC attribute with the highest impact on dual primary). Each internal node represents a test on an attribute (predictive variable: BC attributes in the order of dual primary importance); each branch represents a test output (judgment attribute size); and each leaf node represents a category (result variable: *status*). Data prediction and training start from the root node and proceed step-by-step. Every data point on the non-leaf node is divided into two or more child data sets according to the characteristics of the current node attributes; thereafter, it is passed on to the next level node for processing. After arriving at the leaf nodes, the division of data does not need to continue, and the data of the leaf node are the classification of the prediction [18][24].

Ten-fold cross-validation and grid searching are used to optimize the model, and the ten-fold cross-validation trains the samples into ten shares in the training models. One copy of the reservation is retained as the authentication model data, and the remaining nine samples are used for training. Cross-validation is repeated ten times, and each sample is verified once. Then, the results of the ten cycles are averaged. This prevents model over-fitting and improves the generalizability of the model [18].

Grid search presets several parameter combinations for the model, and each group of hyperparameters is evaluated through cross verification. Finally, the optimal parameter is selected to establish the model and achieve the pruning effect [25].

2.4.4 RF model

The RF algorithm uses the bootstrap method to randomly sample N new self-help sample sets and create N regression trees, which addresses the disadvantage of the over-fitting of a single DT [26][27]. RF is a robust classifier for the training and prediction of samples using multiple DTs. Each weak classifier (i.e., single DT) makes a judgment according to its state and finally votes to select the final classification result, as shown in Fig. 1. The model also uses ten-fold cross-validation and grid search to adjust the over-parameters.

2.5 Model evaluation indicators

In machine learning, the accuracy, precision, recall, F1 score, and AUC are commonly used to measure the ability of two-class classifiers to correctly classify the experimental data. The values of the first four indicators are in [0, 1]. The larger the value, the better is the model effect. The formula is as follows [24][28][29]:

$$accuracy = (TP + TN) / (TP + FN + FP + TN), \tag{2}$$

$$precision = TP / (TP + FP), \tag{3}$$

$$recall = TP / (TP + FN), \tag{4}$$

$$F1 = 2 * (precision * recall) / (precision + recall), \tag{5}$$

where TP is the number of correctly predicted positive cases, TN is the number of correctly predicted negative cases, FP is the number of incorrectly predicted positive cases, and FN is the number of incorrectly predicted negative cases. The F1 score, which considers both the precision and recall of the classification model, can be regarded as a weighted average of the model precision and recall [29]**Błąd! Nie można odnaleźć źródła odwołania.**. The value of the AUC is in the range of [0.5, 1]. Positive and negative samples are randomly selected, and the probability that the positive sample value is higher than the negative sample value becomes the AUC value [24]. The larger the value, the better is the classifier performance.

3. Results

3.1 Model prediction process

After data selection and transformation, the model prediction process, shown in Fig. 2, was completed.

3.2 Parameter tuning results under cross-grid technology

The ten-fold cross-validation and grid search adjusts the size of over-parameters according to the DT model indicators (i.e., accuracy, precision, recall, F1 score, and AUC). After multiple selections, the

index range of the split dataset of the DT is [entropy, Gini]; the maximum depth range of the tree is [2,3,4,5,6,7,8]; and the minimum number of split-leaf samples is [2,3,4,8,12,16,20,24]. Exhaustive searching finds the optimal parameter: {criterion: entropy, max_depth: 5, min_samples_split: 16}. Similarly, the RF algorithm splits the dataset index (criterion) using Gini; the maximum depth of each DT (max_depth) is 9; the number of DTs (n_estimators) is 17; and the minimum split sample size of leaves per tree (min_samples_split) is 18.

3.3 Analysis of model results

The information from 7,826 first-episode BC patients (case:control = 1.05:1) was screened in this study. The DT and RF algorithms were used to train the real-world datasets, and the importance scores of the influential dual primary attributes were obtained, as listed in Table 3. Surgical information (Surgery) and lymph-node status (RNP, RNE) considerably influence the occurrence of dual primary BC. The presence of distant metastasis (AJCCM), tumor size (tumorSize), survival time (times), histologic type (Histologic), and age at diagnosis (Age) are the characteristic expressions of the dual primary probability of first-episode BC, which has adequate reference significance for studying the dual primary probability of first-episode BC patients.

The DT model was constructed, and the combined risk-factor variables were compiled. Surgery, RNP, RNE, tumorSize, times, Histologic, and Age were used for node splitting, and multiple paths were output, as shown in Fig. 3.

The comparison results of the model are summarized in Table 4. The accuracy, precision, recall, F1 score, and AUC of the KNN model ranged from 75.85% to 76.50%, which were considerably better than those of the MultinomialNB (62.92% to 64.36%) and LR (67.57% to 69.86%) models. The MultinomialNB model assumes that the attributes are independent of each other, and the classification effect is not sufficient when the attributes are highly correlated. The limitation of the LR model is that it cannot process nonlinear and highly correlated data. Moreover, the screened BC data variables were complicated and dependent on each other; thus, the prediction effect of these models was not ideal. Although the KNN model can be used for nonlinear classification, its ability to deal with highly correlated data is limited, and the prediction result is inferior to those obtained using the other five models (LSTM, RNN, 1D-CNN, DT, and RF).

The blue shadow highlights that RF exhibited the highest accuracy (97.19%), precision (95.01%), recall (99.48%), F1 score (97.19%), and AUC (97.10%), as listed in Table 4. The gray-shaded indicator value is second only to the RF indicator value. It can be inferred that the performances of the 1D-CNN, LSTM, and DT models were adequate, although not comparable to that of RF. LSTM and 1D-CNN are deep learning models, which mainly process image data (at least three dimensions) and require a large number of datasets of different instances. However, the initial data in this study were two-dimensional

tabular data, which could be used only after raising the dimension through the *expand_dims* function, and the amount of data was limited. The DT algorithm is easy to fit a single tree. Hence, the robust classifier RF algorithm comprising multiple DTs had the best effect. The receiver operating characteristic curves of the eight machine learning algorithms are compared in Fig. 4. The RF curve is closest to the upper left corner and has the best performance.

3.4 External verification

To verify the portability and generalization of the RF model, we used the BC data of the SEER database from 2005 to 2009 for external verification. As HER2 information for the period before 2010 was missing, the externally validated dataset had only 18 predictive attributes. The datasets underwent preprocessing steps such as numerical replacement and missing value processing. There were 202,018 (approximately 98.53%) and 3,012 (approximately 1.47%) cases in the control and case groups, respectively. Similarly, we extracted 3000 control data using the undersampling method, which brought the control:case ratio close to 1:1, to ensure balance between the two groups of data. Then, the total amount of data was reduced to 6,012 (3000 + 3012). The *joblib.dump* function was used to save the eight models after internal validation, and the *joblib.load* function was used to locally call them back. The evaluation results of the models under external verification are listed in Table 5. The RF model was still the best as per all indicators, with the highest accuracy (97.34%), precision (96.67%), recall (97.97%), F1 score (97.32%), and AUC (97.34%). Moreover, RF was the only model whose AUC increased (0.24%) under external verification, as summarized in Supplementary Table S2. These results suitably demonstrate that the RF model exhibits adequate portability and generalization.

4. Discussion

In this study, we used eight machine learning methods (i.e., MultinomialNB, LR, KNN, LSTM, RNN, 1D-CNN, DT, and RF) to predict dual primary cancer in first-episode BC patients, whose data were obtained from the SEER database. The model evaluation index value was high, which can effectively assist doctors in diagnosis and treatment. With the participation of first-episode BC patients in treatment, physicians can select the prediction model with the best evaluation performance. By considering the visualized output results of the model, physicians can input the specific attribute field values of patients (experimental influence factors of dual primary disease) according to the order of their importance; furthermore, they can obtain the output results to predict whether the patient will develop a dual primary disease. However, the actual surgical treatment and follow-up plan are based on the experience of physicians and the actual condition of patients, and the results of the study can serve only as an auxiliary reference.

We adopted a combination of multiple interpolations and machine learning to process missing data and present the model data relationships. The Alma research team in the United States observed that when the missing data were continuous or categorical variables, the missing rate was low (<30%) [30]. The use of multiple interpolation methods was suggested, because such methods are simple, convenient, and easy to operate, with less influence on the analysis results. After cleaning and transformation of the data in this study, there were still some missing values, including 270 missing data (approximately 0.36%) in the control group and 345 missing data (approximately 0.47%) in the case group. To ensure the authenticity of the samples, multiple interpolation methods were used to fill the missing values. In feature engineering, multiple interpolation is used to deal with missing data, which can provide a reference for machine learning engineers.

The model evaluation in this study measured the classification ability of the machine learning methods based on the accuracy, precision, recall, F1 score, and AUC. The prediction performance of the RF model was better than those of the deep learning (1D-CNN, RNN, and LSTM) models, with the highest accuracy (97.19%), precision (95.01%), recall (99.48%), F1 score (97.19%), AUC (97.10%); the external verification also yielded the same outcome. Furthermore, our prediction results are superior to those of image processing based on contemporary deep learning methods. For instance, Mishra et al. used a deep neural network to preprocess, segment, and classify thermal images, and the prediction accuracy of the output spectrum of 680 heat map training data for BC reached 0.958, which was lower than 0.97 achieved by the RF model in this study [31]. Devi et al. applied a deep neural network to intelligent image analysis and found that deep learning could diagnose various cancers, including cervical, breast, colon, and lung, with the highest accuracy of 0.92, which was again lower than 0.97 achieved by the RF model in this study [32].

Early studies on dual primary BC were mainly focused on the summary of clinicopathological practices and genome detection [5][8]. Referring to other machine learning-based BC prediction cases [33][35], this study applied machine learning to the dual primary prediction of BC for the first time. Our method can aid doctors in the diagnosis and treatment of first-episode BC with theoretical and medical significance. Currently, the development of dual primary cancers can be predicted based on only reviews, predictive screenings, and physicians' decisions. Physicians' decision-making based on the TNM staging system is currently the most widely used method for evaluating the prognosis. However, it has some limitations, especially for patients with dual primary malignancies (these tend to have special biological characteristics different from those of single primary malignancy) [36]. In addition, its high workload, longer analysis time, subjective biases, and insufficient data may lead to over- or under-treatment. In contrast, when compared with the traditional methods, RF can predict quickly and accurately, and its predictive value is considered superior to those of other evaluation systems [37][38]. Considering the

good prediction performance and clinical utility of this RF model, it is expected to be routinely applied to dual primary prediction of BC patients in the future.

Although our study performed well in identifying dual primary BC, it may have some limitations. The potential correlation between BC and other cancers was not considered in the prediction of dual primary cancer; for example, the probability of dual primary cancer may be higher in first-episode BC patients with thyroid cancer (TC) [39]. This requires the identification of all possible related types of cancer under a physician's guidance by using prognostic factors to estimate their associations and to explore their commonalities and differences. Zhang et al. found that the overall risk of the occurrence of a second primary TC or BC increased in patients with BC or TC. TC and BC may have a strong association between their primary mechanisms [40]. We have observed that thyroid and BC prognosis factors may have an association. A comparative study will be conducted between patients with first-episode BC alone and patients with TC combined with first-episode BC to explore the probability and influencing factors of dual primary BC. We will also study whether TC affects the prognosis of dual primary BC patients. Additionally, to address the problem of unbalanced classification, this study adopted an undersampling method, but it could adjust the imbalance only to a limited degree [41]. Therefore, our future research will deal with the problem of unbalanced classification.

The findings of this study demonstrate that distant metastasis attributes are influential factors in the prediction of dual primary tumors in first-episode BC patients. However, the prediction of dual primary cancer and metastatic cancer is an important research issue in practical medical scenarios, and it is difficult for doctors to distinguish between the two. Doctors must formulate different treatment plans for different cancers. Liu et al. reported that adenoid cystic carcinoma in the breast rarely spreads via the lymphatic system and mainly affects the visceral organs, with the most frequent site of metastasis being the lung [42]. Therefore, our subsequent research will focus on using machine learning methods to build a model that determines whether the regenerative malignancy of first-episode BC patients is metastatic or dual primary cancer, considering the potential correlation between BC and other cancers.

Acknowledgments and Funding Information

This study was supported by the Natural Science Research Project of Universities in Anhui Province in 2017 (key) "Endocrine personalized therapy for breast cancer patients based on Data Mining technology" (KJ2017A390). We would like to thank Editage (www.editage.cn) for English language editing.

References:

[1] Senkus E, Kyriakides S, Penaultllorca F, et al. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Saunders. Annals of oncology 2015; 26: v8-30.

[2] Kim H, Lee H, Choi DH, et al. Distribution of tumor subtypes in bilateral breast cancer: Comparison between synchronous and metachronous cancer. Asia-Pacific journal of clinical oncology 2020; 7:1-8.

[3] Giannakeas V, Lim DW, Narod SA. The risk of contralateral breast cancer: a SEER-based analysis. Br J Cancer. 2021; 125:601-610.

[4] Wei JL, Jiang YZ, Shao ZM. Survival and chemotherapy-related risk of second primary malignancy in breast cancer patients: a SEER-based study. International journal of clinical oncology 2019; 24: 934-940.

[5] Bo J, Zhang S, Xin C, et al. Breast cancer and synchronous multiple primary lung adenocarcinomas with heterogeneous mutations: a case report. BMC Cancer 2018; 18: 1138.

[6] Mruthyunjayappa S, Zhang K, Zhang L, Eltoum IA, Siegal GP, Wei S. Synchronous and metachronous bilateral breast cancer: clinicopathologic characteristics and prognostic outcomes. Hum Pathol. 2019; 92: 1-9.

[7] Jena A, Patnayak R, Lakshmi AY, Manilal B, Reddy MK. Multiple primary cancers: An enigma. South Asian J Cancer. 2016; 5: 29-32.

[8] Kitada, K, Sugi, K, Matsuoka T, et al. A case of solitary pulmonary metastasis of breast cancer intraoperatively diagnosed as primary lung cancer 36 years after the first surgery. Nihon Rinsho Geka Gakkai Zasshi (Journal of Japan Surgical Association) 2005; 66: 1303–1307.

[9] Criscitiello C, Azim HA Jr, Agbor-tarh D, et al. Factors associated with surgical management following neoadjuvant therapy in patients with primary HER2-positive breast cancer: results from the NeoALTTO phase III trial. Ann Oncol. 2013; 24: 1980-5.

[10] Beckmann KR, Buckingham J, Craft P, et al. Clinical characteristics and outcomes of bilateral breast cancer in an Australian cohort. Breast. 2011; 20: 158-64.

[11] Wu J, Hicks C. Breast Cancer Type Classification Using Machine Learning. J Pers Med. 2021; 11:61.

[12] Qin J, Deng G, Ning J, Yuan A, Shen Y. Estrogen Receptor Expression on Breast Cancer Patients' Survival under Shape Restricted Cox Regression Model. Ann Appl Stat. 2021; 15:1291-1307.

[13] Teichgraeber DC, Guirguis MS, Whitman GJ. Breast Cancer Staging: Updates in the AJCC Cancer Staging Manual, 8th Edition, and Current Challenges for Radiologists, From the AJR Special Series on Cancer Staging. AJR Am J Roentgenol. 2021; 217:278-290.

[14] Dobashi N, Saito S, Nakahara Y, Matsushima T. Meta-Tree Random Forest: Probabilistic Data-Generative Model and Bayes Optimal Prediction. Entropy (Basel). 2021; 23: 768-786.

[15] Daly MC, Paquette IM. Surveillance, Epidemiology, and End Results (SEER) and SEER-Medicare Databases: Use in Clinical Research for Improving Colorectal Cancer Outcomes. Clin Colon Rectal Surg. 2019; 32:61-68.

[16] Lloyd MR, Stephens SJ, Hong JC, et al. The impact of COVID-19 on breast cancer stage at diagnosis. Journal of Clinical Oncology 2021; 39: 528-528.

[17] Pan Y, Gao H, Lin H, Liu Z, Tang L, Li S. Identification of Bacteriophage Virion Proteins Using Multinomial Naïve Bayes with g-Gap Feature Tree. Int J Mol Sci. 2018; 19:1779.

[18] Chao CM, Yu YW, Cheng BW, Kuo YL. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. J Med Syst. 2014; 38:106.

[19] Zakaria L, Ebeid HM, Dahshan S, et al. Analysis of Classification Methods for Gene Expression Data. In: Springer, Cham. 2020: 19:190-199.

[20] Tahir M, Hayat M, Kabir M. Sequence based predictor for discrimination of enhancer and their types by applying general form of Chou's trinucleotide composition. Comput Methods Programs Biomed. 2017; 146:69-75.

[21] Sanderson M, Bulloch AG, Wang J, Williamson T, Patten SB. Predicting death by suicide using administrative health care system data: Can recurrent neural network, one-dimensional convolutional neural network, and gradient boosted trees models improve prediction performance? J Affect Disord. 2020; 264:107-114.

[22] Liu M, He L, Hu B, Li S. Recurrent neural network with noise rejection for cyclic motion generation of robotic manipulators. Neural Netw. 2021; 138:164-178.

[23] Mumtaz W, Qayyum A. A deep learning framework for automatic diagnosis of unipolar depression. Int J Med Inform. 2019; 132:103983.

[24] Lee JY, Lee KS, Seo BK, et al. Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on MRI. European radiology 2021; 32: 1-11.

[25] Zhou H, Zhong Z, Hu M, Huang J. Determining the steering direction in critical situations: A decision tree-based method. Traffic Inj Prev. 2020; 21:395-400.

[26] Gutiérrez-Cárdenas J,Wang Z. Classification of Breast Cancer and Breast Neoplasm Scenarios Based on Machine Learning and Sequence Features from lncRNAs-miRNAs-Diseases Associations. Interdiscip Sci 2021; 13: 572-581.

[27] Macaulay BO, Aribisala BS, Akande SA, Akinnuwesi BA, Olabanjo OA. Breast cancer risk prediction in African women using Random Forest Classifier. Cancer Treat Res Commun. 2021;

28:100396.

[28] Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019; 19:281.

[29] DeVries Z, Locke E, Hoda M, et al. Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability. Spine J. 2021; 21:1135-1142.

[30] Alma P, Ellen M, Deirdre CF, et al. Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology 2017; 9: 157-166.

[31] Mishra S, Prakash A, Roy S K, et al. Breast cancer detection using thermal images and deep learning. IEEE 2020: 211-216.

[32] Devi VA, Nayyar A. Fusion of deep learning and image processing techniques for breast cancer diagnosis. Springer, Singapore 2021: 1-25.

[33] Wang B, Yuan F. The association between Interleukin-1β gene polymorphisms and the risk of breast cancer: a systematic review and meta-analysis. Archives of Medical Science. 2022; 18: 1-10.

[34] Shen Y, Peng X, Shen C. Identification and validation of immune-related lncRNA prognostic signature for breast cancer. Genomics. 2020; 112: 2640-2646.

[35] Paik HJ, Jung YJ, Kim DI, et al. Clinicopathological Features of BRCA1/2 Mutation-Positive Breast Cancer. Oncology. 2021; 99:499-506.

[36] Song C, Yu D, Wang Y, et al. Dual Primary Cancer Patients With Lung Cancer as a Second Primary Malignancy: A Population-Based Study. Front Oncol. 2020; 10: 515606.

[37] Ning S, Li H, Qiao K, Wang Q, et al. Identification of long-term survival-associated gene in breast cancer. Aging (Albany NY). 2020; 12: 20332-20349.

[38] Toth R, Schiffmann H, Hube-Magg C, et al. Random forest-based modelling to detect biomarkers for prostate cancer progression. Clin Epigenetics. 2019; 11: 148.

[39] Dong L, Lu J, Zhao B, Wang W, Zhao Y. Review of the possible association between thyroid and breast carcinoma. World J Surg Oncol. 2018; 16:130.

[40] Zhang L, Wu Y, Liu F, Fu L, Tong Z. Characteristics and survival of patients with metachronous or synchronous double primary malignancies: breast and thyroid cancer. Oncotarget. 2016; 7: 52450-52459.

[41] Bria A, Marrocco C, Tortorella F. Addressing class imbalance in deep learning for small lesion detection on medical images. Comput Biol Med. 2020; 120: 103735.

[42] Liu Z, Wang M, Wang Y, Shen X, Li C. Diagnosis of adenoid cystic carcinoma in the breast: a case report and literature review. Archives of Medical Science. 2022 ;18: 279-283.

Table 1 Selected fields (categorical)

Categorical variables name	Shorthand	Explain	Number of categories
Race recode (White, Black, Other)	**Race**	**race**	3
Marital status at diagnosis	**Marital**	**Marital status**	2
Primary Site	**primarySite**	**Site of primary lesion**	10
Laterality	**Laterality**	**Unilateral/bilateral (breast cancer)**	3
Histologic Type ICD-O-3	**Histologic**	**Histological type**	43
Grade	**Grade**	**Histological grading**	3
CS lymph nodes	**lymphNodes**	**Lymph node points**	31
Regional nodes positive	**RNP**	**Region node positive**	38
Regional nodes examined	**RNE**	**Region node examined**	50
Derived AJCC T, 7th ed	**AJCCT**	**Tumor size**	5
Derived AJCC N, 7th ed	**AJCCN**	**Regional lymph node involvement**	4
Derived AJCC M, 7th ed	**AJCCM**	**presence of distant metastasis**	2
ER Status Recode Breast Cancer	**ER**	**Estrogen status**	2
PR Status Recode Breast Cancer	**PR**	**Progesterone state**	2
Derived HER2 Recode	**HER2**	**HER2(Human epidermal growth factor receptor) status**	2
RX Summ--Surg Prim Site	**Surgery**	**surgical information**	2

Table 2. Selected fields (continuous)

Continuous variables names	Shorthand	Explain	Range
Age at diagnosis	**Age**	**age at diagnosis**	**20~90**
CS tumor size	**tumorSize**	**tumor size**	**0~998**
Survival months	**times**	**survival time**	**0~83**

Table 3. Attribute importance score table.

Variables	Decision tree（%）	Random forest（%）
Race	**<0.001**	**0.135154**
Marital	**<0.001**	**0.058861**
primarySite	**<0.001**	**0.356892**
Laterality	**<0.001**	**0.138930**
Histologic	**0.223084**	**0.509991**
Grade	**<0.001**	**0.381362**
lymphNodes	**<0.001**	**0.625182**
RNP	**8.702041**	**1.530730**
RNE	**0.555700**	**1.710177**
AJCCT	**<0.001**	**0.306317**
AJCCN	**<0.001**	**0.208543**
AJCCM	**1.304067**	**0.843754**
ER	**<0.001**	**0.085416**
PR	**<0.001**	**0.173831**
HER2	**<0.001**	**0.148102**
Surgery	**88.135351**	**80.247779**
Age	**0.196668**	**2.287591**
tumorSize	**0.564057**	**0.717006**
times	**0.319027**	**9.534372**

Table 4. Comparison of machine learning classification performance.

Model name	Accuracy	Precision	Recall	F1 score	AUC
MultinomialNB	**0.6325**	**0.6292**	**0.6436**	**0.6363**	**0.6325**
LR	**0.6887**	**0.6757**	**0.6986**	**0.6870**	**0.6800**
KNN	**0.7585**	**0.7650**	**0.7631**	**0.7640**	**0.7598**
1D-CNN	**0.9655**	**0.9425**	**0.9895**	**0.9654**	**0.9660**
RNN	**0.9621**	**0.9450**	**0.9801**	**0.9622**	**0.9624**
LSTM	**0.9625**	**0.9492**	**0.9768**	**0.9628**	**0.9628**
DT	**0.9659**	**0.9465**	**0.9861**	**0.9659**	**0.9664**
RF	**0.9719**	**0.9501**	**0.9948**	**0.9719**	**0.9710**

Table 5. Comparison of model performance under external validation.

Model name	Accuracy	Precision	Recall	F1 score	AUC
MultinomialNB	**0.6286**	**0.6267**	**0.6281**	**0.6274**	**0.6269**
LR	**0.6824**	**0.6778**	**0.6831**	**0.6804**	**0.6785**
KNN	**0.7539**	**0.7556**	**0.7522**	**0.7539**	**0.7532**
1D-CNN	**0.9590**	**0.9378**	**0.9791**	**0.9580**	**0.9589**
RNN	**0.9573**	**0.9544**	**0.9598**	**0.9571**	**0.9573**
LSTM	**0.9618**	**0.9567**	**0.9563**	**0.9565**	**0.9617**
DT	**0.9645**	**0.9633**	**0.9655**	**0.9644**	**0.9645**
RF	**0.9734**	**0.9667**	**0.9797**	**0.9732**	**0.9734**

Table S1. List of abbreviations.

Abbreviation	Explanation
NCI	**National Cancer Institute**
BC	**breast cancer**
BBC	**bilateral breast cancer**
MBBC	**metachronous bilateral breast cancer**
SBBC	**simultaneous bilateral breast cancer**
UBC	**unilateral breast cancer**
TC	**thyroid cancer**
SEER	**The Surveillance, Epidemiology, and End Results**
LR	**logistic regression**
KNN	**K-nearest neighbor**
BN	**Naive Bayes**
DT	**decision tree**
RF	**random forest**
1D-CNN	**one-dimensional convolutional neural network**
RNN	**recurrent neural network**
LSTM	**long short-term memory**
ROC	**receiver operating characteristic**
AUC	**area under the receiver operating characteristic curve**

Table S2. Internal and external validation differences.

Indicators	Multino-mialNB	LR	KNN	1D-CNN	RNN	LSTM	DT	RF
Internal AUC_1	**0.6325**	**0.6889**	**0.9495**	**0.9660**	**0.9624**	**0.9628**	**0.9664**	**0.9710**
External AUC_2	**0.6269**	**0.6880**	**0.9540**	**0.9589**	**0.9573**	**0.9617**	**0.9645**	**0.9734**
AUC_2-AUC_1	**-0.0056**	**-0.0009**	**-0.0045**	**-0.0071**	**-0.0051**	**-0.0011**	**-0.0019**	**0.0024**

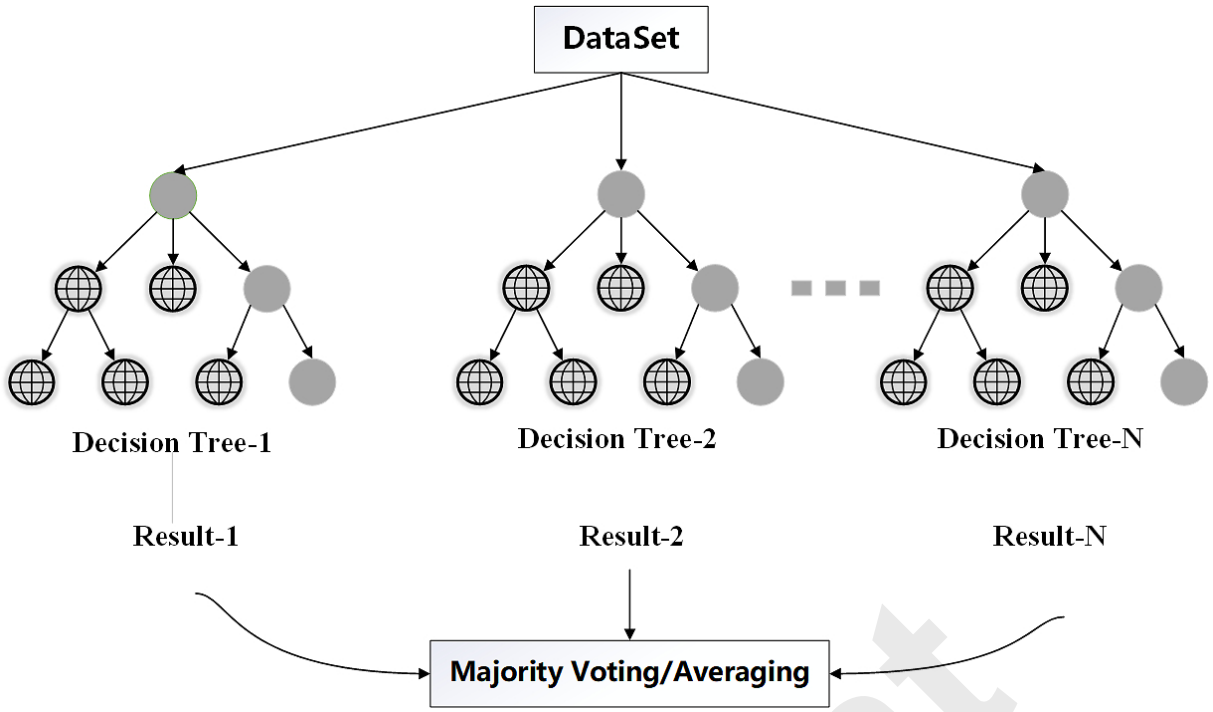

Fig. 1. Random forest structure.

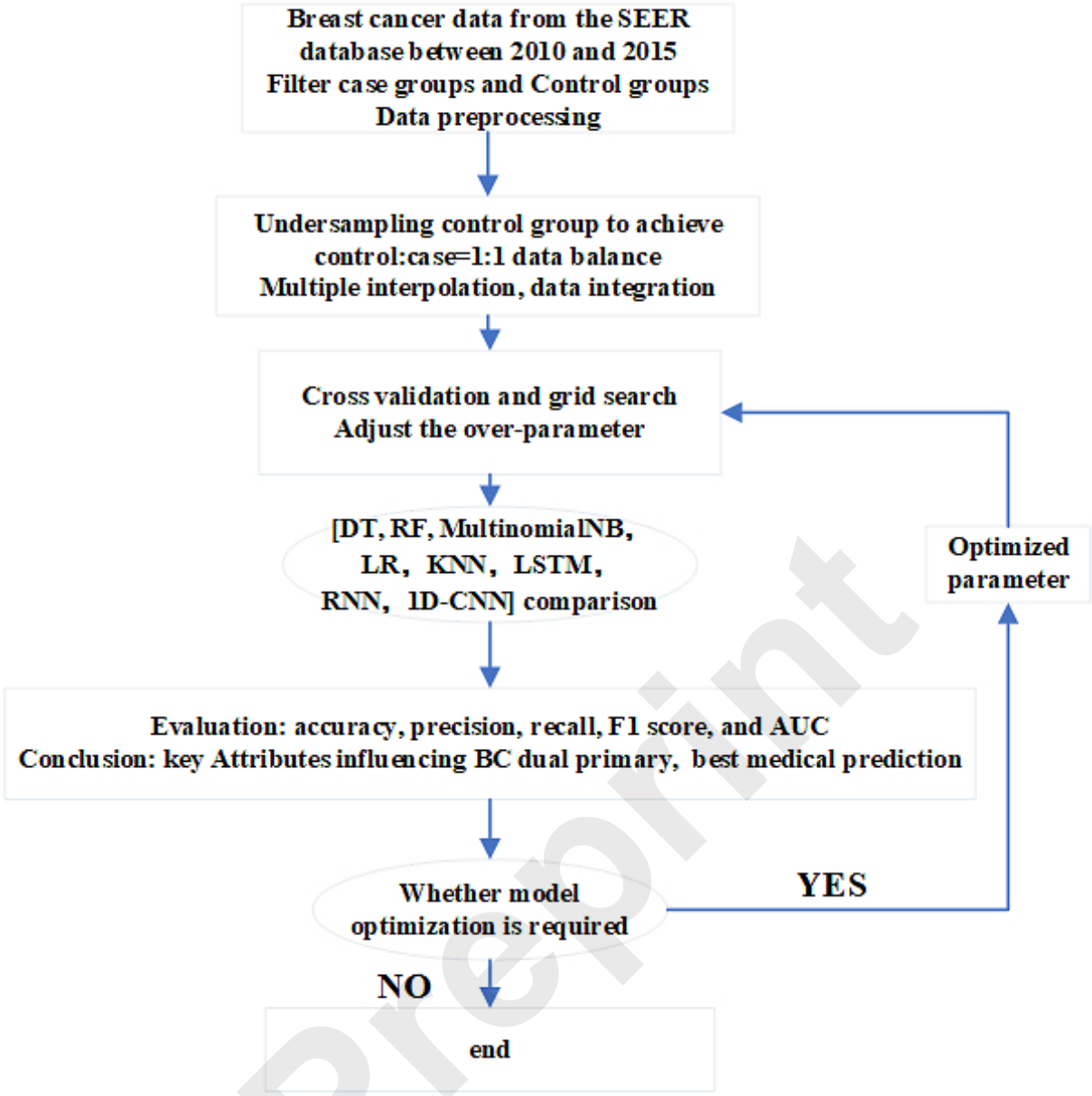

Fig. 2. Model prediction process.

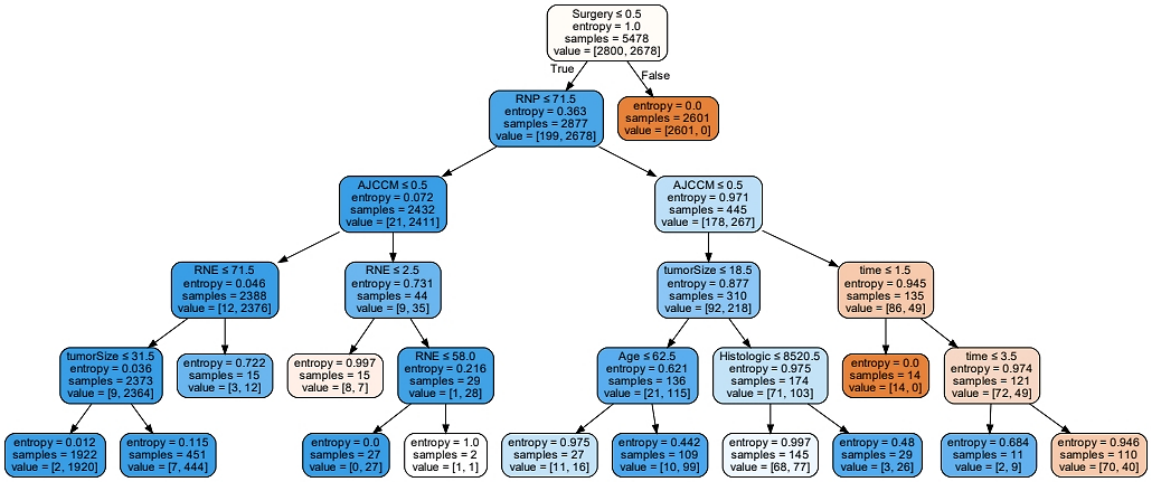

Fig. 3. Visualization result of decision tree.

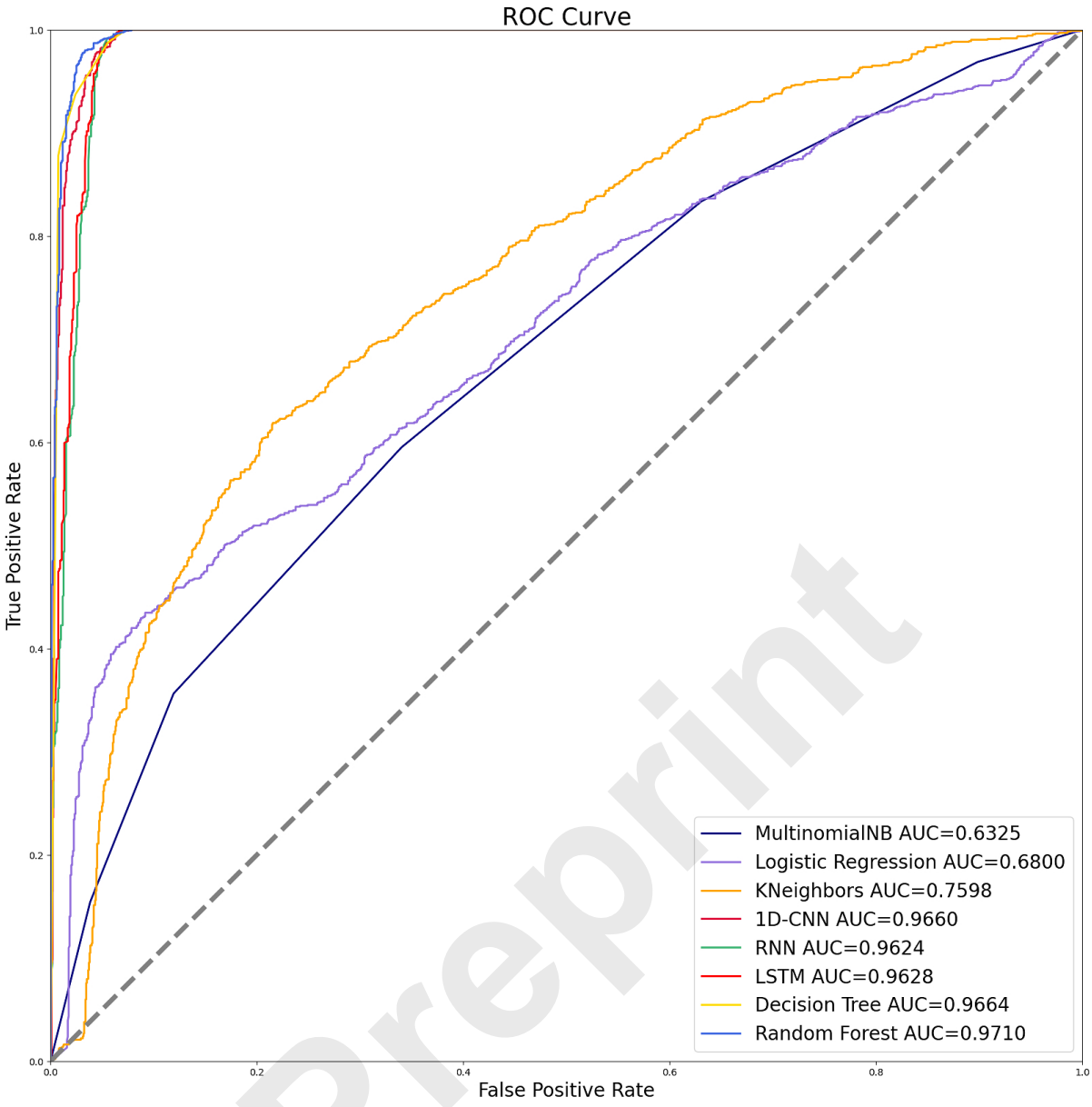

Fig. 4. ROC curve comparison of machine learning models.