Multi-label, Classification-based Prediction of Breast Cancer Metastasis Directions

Tingting Wang, Qi Fan, Liang Tan and Beier Zhang

Abstract—**Predicting the metastatic direction of primary breast cancer (BC), thus assisting physicians in precise treatment, strict follow-up, and effectively improving the prognosis. The clinical data of 293,946 patients with primary BC diagnosed between 2010 and 2015 were collected from the Surveillance, Epidemiology, and End Results database. Multiple interpolations and Multi-label Synthetic Minority Over-sampling Technique methods were used for data analysis, and machine learning model was established for multi-label classification. Finally, Surgical information, lymph node status, distant metastasis, tumor size, chemotherapy, histological type, and radiotherapy had significant influence as inputs. Compared with the k-nearest neighbor model, average accuracies of the decision tree and random forest (RF) models increased from 88.84% to 93.59% and 94.14%, respectively. Their average precision, recall rate, F1 score, area under the receiver operating characteristic curve and weighted-F1 increased from 87.24% to 95.85% and 94.74%, 87.73% to 90.40% and 91.76%, 87.07% to 92.16% and 93.45%, 97.11% to 99.53% and 99.95%, 82.13% to 89.44% and 90.48%, respectively. In conclusion, the RF model, which showed the best performance, can be used in multi-label prediction of BC metastasis directions, and can assist physicians in diagnosing and treating patients with primary BC.**

Index Terms—**Breast cancer, multi-label, metastasis directions, prediction, decision tree, random forest.**

I. Introduction

WITH advances in oncology medicine, the mortality rate in cancer patients has decreased; however, during follow-up, the rate of cancer metastasis in surviving cancer patients is increasing [1]. Metastatic breast cancer (MBC) is a heterogeneous disease. It has a variety of clinical manifestations, ranging from isolated metastases to diffuse and multi-organ involvement [2]. As reported, cancer metastasis rate in BC patients after diagnosis and rate of primary tumor treatment are as high as 20-30%, and approximately 90% of cancer-related deaths are attributed to metastasis [3]. After the first treatment of BC patients, the risk of metastasis has an adverse effect on the patients, and is an important prognostic factor. In patients with primary BC (cured either by partial mastectomy, modified radical mastectomy, radiotherapy, or chemotherapy), the average time to organ metastasis is 3.7 years after diagnosis, and the sites of metastasis are likely to the bones (39.80%), lung (10.94%), liver (7.34%), or brain (1.51%) [4]. However, the possibility of recovery after the occurrence of MBC is extremely low, and the 5-year survival rate is reduced from $>80\%$ [5] (according to the original "global surveillance of cancer survival" [In 2015, the second cycle of the Global Cancer Survival Surveillance could serve as a metric of the effectiveness of health systems and inform global cancer control policy]) to approximately 25% [6]. Therefore, it is crucial to determine whether organ metastasis occurs in patients diagnosed with primary BC. If the direction of metastasis can be accurately predicted, targeted treatment can be carried out; Then, strict review of patients to avoid improper treatment, in order to improve the prognosis of patients.

Puppo et al. discussed the possibility of using miRNAs as direct therapeutic targets or advanced therapies for BC bone metastasis; also for their potential as predictive biomarkers of bone metastasis for early diagnosis and better tailoring of therapies for patients with cancer [7]. Studies have demonstrated that serum miRNA profiling may serve as a biomarker for eribulin responsiveness [A synthetic field soft sponge analogue produced by Wei Cai, Japan, has a novel mechanism of action and is approved for the treatment of BC and liposarcoma] and for predicting the development of new distant metastases in MBC [8]. Feng et al. used a lentivirus vector-based shRNA technique to test the functional relevance of cellular retinoic acid binding protein 2 (CRABP2) knockdown in breast tumors [9]. They demonstrated that CRABP2 could inhibit the invasion and metastasis of estrogen receptor-positive (ER+) BC by regulating the stability of Lats1 in vitro and in vivo; and subsequently promote the invasion and metastasis of estrogen receptor-negative (ER-) BC, which provides a new idea for BC treatment. Early research on MBC has mainly focused on innovations in gene detection technology. Gene detection is used to detect mutations in tumor cells. Biopsy or histopathological section samples should be obtained. However, puncture biopsy has some disadvantages, such as large surgical trauma, and is unsuitable in patients with surgical contraindications [10]. As histopathological sections are obtained during surgery; they are only suitable for intraoperative and postoperative (formalin-soaked for 1-2 years) detection, not for preoperative detection or long-term postoperative follow-up (2 years) [10]. Although the recently popularized "liquid biopsy" gene detection method (the spread of advanced cancer cells to the blood, which can be captured by gene detection) is

Manuscript received April 8, 2024; revised October 27, 2024. This study was supported by the Anhui Provincial Quality Engineering Project "Construction and Practice of New Computer Science and Technology Specialities under the Background of Industry-Teaching Integration" (2023sx154) and the Anhui Province New Era Education Quality Project "Application of Double Weighted Random Forest Multi-Label Algorithm in Predicting the direction of breast cancer Metastasis" (2022xscx090). It was also 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).

Tingting Wang a teacher of the School of Computer and Software Engineering, Anhui Institute of Information Technology, China (e-mail: 2128296365@qq.com).

Qi Fan is an associate professor of School of Computer Science and Technology, Huaibei Normal University, China (corresponding author to provide phone: 13966139306; e-mail: 8073592@qq.com).

Tan Liang is an associate professor at the School of Computer and Software Engineering, Anhui Institute of Information Technology, China (e-mail: 287500336@qq.com).

Beier Zhang is a postgraduate student of School of Computer Science and Technology, Huaibei Normal University, China (e-mail: 2530066750@qq.com).

non-invasive and of low-risk; however, it only produces a small number of ctDNA targets, induce false-positive and false-negative results, and is mostly suitable for patients with advanced cancer [11]. Therefore, a new detection technology is urgently needed to replace such traumatic and non-universal examinations, to provide a good reference for patients' clinical treatment and recovery of prognosis.

Holm et al. used multiple logistic regression analyses to determine the outcome (ER status and lymph node involvement) [12]; The Cox proportional risk model was also used to estimate the risk ratio of distant metastasis. Cheng et al. used the k-nearest neighbor (KNN) and selection operator regression algorithm to train filtered and normalized lattice radiomic features, to enable patients achieve better early stratification, brain metastasis screening, and overall prognosis [13]. Yang et al. constructed a nomogram for liver metastasis based on multivariate logistic regression analysis to facilitate the preventive treatment or monitoring of liver metastasis [14]. The use of machine learning regression models to predict the prognosis of BC metastasis is a common clinical analysis method; however, such regression models cannot process non-linear and highly correlated data. Moreover, in clinical medical data, it is usually difficult for attribute variables to be independent of one another. For example, the primary tumor size (derived from the 7th edition of the American Joint Committee on Cancer [AJCC T, 7th Ed]), regional lymph node involvement (from AJCC N, 7th Ed), and presence or absence of distant metastasis (AJCC M, 7th Ed) all influence one another when determining tumor staging [15], [16]. However, the decision tree (DT) and random forest (RF) algorithms can effectively process collinear variables and indirectly improve the accuracy and recall rate of a model. Therefore, this type of algorithm is suitable for cancer prediction research, in which variables interact with each other.

Mercan et al. used four different multi-instance and multi-label learning algorithms to perform sliding-level and roy level predictions; In order to diagnose the pathological images of breast tissue, multi-class localization can be realized [17]. Qu et al. implemented a multi-criterion mammographic risk analysis system using multi-label fuzzy-rough feature selection [18]. In practical applications, samples often contain multiple tags. For example, a BC patient may develop organ metastases in different directions after an initial treatment; it is therefore inaccurate to classify the metastases as one type. Therefore, to analyze the pathological data more accurately, the classification of BC organ metastasis should be described as a multi-label problem; that is, each point in the training set is associated with multiple labels. However, not all traditional machine learning algorithms are suitable for multi-label classifications [19]. Only KNN, DT, RF, et al. support multi-label classification. Because genes could be associated with multiple molecular functions, Fodeh et al. suggested that the gene ontology molecular function annotation is a multi-label classification problem with several classes; therefore, they used the KNN classifier to carry out classification experiments, which performed better [20]. Studies demonstrate that DT can capture the relations between labels and analyze a set of rules to treat multi-label problems [19]. Zhou et al. used the RF model to predict four diabetic complications simultaneously (multi-label classification problem) [21]. Therefore, in this study, the KNN, DT, and RF algorithms were used to build a medical prediction model, expand the multi-label classification, and determine the factors with the greatest influence on the direction of BC metastasis, to provide a theoretical basis for clinicians' diagnosis and treatment.

II. MATERIALS AND METHODS

A. Data collection

The Surveillance, Epidemiology, and End Results (SEER) program, a clinical database funded by the National Cancer Institute, collects data on cancer incidence and survival from U.S. cancer registries [22]. In recent years, the use of machine learning and statistical methods to study the prognosis of cancer information based on the SEER database has become an important medical auxiliary approach.

Clinical data of 293,946 patients with primary BC were collected from SEER database and organized on BC metastasis from 2010 to 2015. The inclusion criteria were as follows: (1) female sex, (2) aged 20 to 80 years at diagnosis, (3) diagnosed between 2010 and 2015, (4) primary BC, (5) single tumor. Exclusion criteria were as follows: (1) diagnosis by autopsy or using a death certificate, (2) survival record of 0 or unknown, (3) incomplete clinicopathological data.

Variable selection: Independent variables included 21 clinicopathological features. Categorical and continuous variables are presented in Tables I and II, respectively. The dependent variables were MetsTotal, MetsBone, MetsLung, MetsLiver, and MetsBrain. These six fields were dichotomous variables, and their attribute characteristics are shown in Table III.

B. Data preparation

Data preprocessing was to simplify the data to meet the modeling requirements as much as possible. First, of an overall data size of 7,642,596 (293,946 * [21 + 5]), 161,695 (approximately 2.12%) were missing. Low loss rate (<30%), it is recommended to use multiple interpolation methods. Because these methods are simple, convenient, and easy to operate, they have little impact on the analysis results [23], [24]. To maintain the authenticity of the samples, missing values were imputed by multiple interpolations.

Surgical information (RX Summ-Surg Prim Site, Surg-Prim) initially had 47 categories, which were too detailed and insufficiently representative. These were then divided into dichotomous variables (SurgPrim=1, surgery; SurgPrim=1, no surgery).

The initial age at diagnosis (Age recoded with <1-year olds) ranged from 1 to 103 years. However, in terms of clinical practice, patients aged 1-20 years have a low probability (<0.1%) and a good prognosis; whereas patients aged 81-103 years are less likely to be cured and have a low representation. Considering the data balance and applicability of the model, only those aged 20 to 80 years were retained. The primary tumor size (derived AJCC T, 7th Ed), a categorical variable, includes T0, T1, T2, T3, and T4 categories, at increasing levels of severity [25]. BC data also include smaller categories, such as T1 subdivided into T1a, T1b, T1c, T1mic, and T1NOS. To reduce the complexity of the

TABLE I
SELECTION FIELDS OF INDEPENDENT VARIABLES (CATEGORICAL)

Categorical variables name	Short form	Definitions	Number of categories
Race recode (White, Black, Other)	Race	Race	3
Age recoded with <1-year olds	Age	Age at diagnosis	6
Year of diagnosis	YD	Year of diagnosis	6
Marital status at diagnosis	Marital	Marital status	2
Laterality	Laterality	Unilateral/bilateral (breast cancer)	3
Primary Site	PrimarySite	Site of primary lesion	10
Chemotherapy recode (yes, no/unk)	Chemotherapy	Chemotherapy	2
Radiation recoded	Radiation	Radiation	2
Grade	Grade	Histological grading	3
Histologic Type ICD-O-3	Histologic	Histological type	54
CS lymph nodes	LymphNodes	Lymph node points	34
Regional nodes examined	RNE	Region node examined	59
Regional nodes positive	RNP	Region node positive	47
Derived AJCC T, 7th ed	AJCCT	Primary 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
Derived HER2 Recode	HER2	Human epidermal growth factor receptor (HER2) status	2
ER Status Recode Breast Cancer	ER	Estrogen status	2
PR Status Recode Breast Cancer	PR	Progesterone state	2
RX Summ-Surg Prim Site	SurgPrim	Surgical information	2

TABLE II
SELECTION FIELDS OF INDEPENDENT VARIABLES (CONTINUOUS)

Continuous variables names	Short form	Definition	Range
CS tumor size	TumorSize	Tumor size	0-998

TABLE III
DEPENDENT VARIABLE SELECTION FIELDS (CATEGORICAL)

Categorical variables name	Short form	Definition	Number of categories
CS mets at dx (2004-2015)	MetsTotal	Whether metastasis occurred (total)	2
SEER Combined Mets at DX-bone	MetsBone	Whether bone metastases have occurred	2
SEER Combined Mets at DX-lung	MetsLung	Whether lung metastases have occurred	2
SEER Combined Mets at DX-liver	MetsLiver	Whether liver metastases have occurred	2
SEER Combined Mets at DX-brain	MetsBrain	Whether brain metastases have occurred	2

modeling analysis, the study combined the smaller categories and retained only five categories. Similarly, regional lymph-node involvement (derived AJCC, 7th ed) retained N0, N1, N2, and N3 after the combination. The presence or absence of distant metastasis (derived AJCC M, 7th Ed) retained M0 and M1 after the combination.

C. Multi-label Synthetic Minority Over-sampling Technique (MLSMOTE)

In multi-label classification scenarios, the number of distribution instances associated with one class is much lower compared to that of another class, resulting in data imbalance. MLSMOTE can effectively deal with data imbalances in multi-label classification by sampling. The specific steps are as follows [26], [27]:

Step 1: Selection of a few instances

Calculation of the imbalance ratio $LRPL(j)$ for each label:

$$LRPL(j) = \frac{\underset{j'=L_1}{\arg\max}^{L_{|L_1|}}(\sum_{i=1}^{|N|} h(j', Y_i))}{(\sum_{i=1}^{|N|} h(j, Y_i))} h(j, Y_i) = \left(1, j \in Y_i;\ \ 0, j \notin Y_i\right) \tag{1}$$

L and N indicate the number of labels and instances, respectively.

Calculation of the average imbalance rate MIR:

$$MIR = \frac{1}{|L|} \sum_{i=L_1}^{L_{|L|}} LRPL(i) \tag{2}$$

Each label of $LRPL(i) > MIR$ is regarded as a tail label, and the label data are regarded as the minority of the instance data.

Step 2: Generation of feature vectors: with Synthetic Minority Over-sampling Technique (SMOTE), oversampling generates the feature vectors for the new-generation data of the tail tag.

Step 3: The label set is generated by calculating the frequency of occurrence of each label in the reference and neighbouring data points using the ranking method. When the frequency of the label appearance exceeds half of the considered instances, it is considered to fall into the target label set.

D. Improved parameterisation methods

1) Traditional cross-grid methods: The traditional cross-grid consists of two steps: cross-validation and grid search. First k-fold cross-validation divides the transfer direction data into training and validation sample sets, where the training samples are divided into k-1 equal parts and only 1 validation sample is retained. The cross-validation step is repeated k times using the evaluation metrics as a measure to ensure that each sample is sampled and validated once; and the mean value of the k-fold validation results is calculated to estimate the predictive ability of the model. Repeating the validation step multiple times can effectively improve the model generalisation ability and avoid overfitting. The grid search requires manual setting of the parameter dictionary, cyclically calling the cross-validation method to evaluate the model parameters, and finally selecting the optimal parameters to create the model to achieve the pruning effect. k-fold cross-grid method can be used for KNN, DT, and RF to adjust the hyper-parameters.

2) Improved Fully Automated Cross Grid Methods: The traditional cross-grid method requires technicians to repeatedly adjust the interval of the parameter array, move left and right tentatively to observe the evaluation results, and search for the model hyperparameters. However, in practical application, it is difficult for medical practitioners or novice technicians to control the interval adjustment rule, and repeated adjustments will delay medical treatment and research and development time, so we consider to increase the fully-automatic features of the algorithm. The pseudo-code of the improved fully-automatic cross-network algorithm is shown in Algorithm I: firstly, steps 2-4 are added on the basis of the traditional algorithm, and the outer loop and flag markers are added to allow the algorithm to repeatedly perform the evaluation operation when flag==1. Adding steps 9-19 includes two if conditions that keep the previous auc1 evaluation result and the current loop auc2 for comparison; and once auc2>auc1 occurs, change the flag flag←1 and update auc1←auc2 to adjust the parameter ranges and go to the next loop. If auc is reduced or unchanged then flag==0, jump out of the loop and end the algorithm. It should be noted that the improved fully automated cross-grid algorithm only adds fully automated features to avoid manual parameter tuning and does not change the parameter selection or evaluation results of the model. The KNN, DT and RF models in the study are all adjusted for hyperparameters by the fully automatic 10-fold cross-grid, which can effectively improve the efficiency of model operation.

Algorithm I: Improved fully automated cross grid

```
Input: dictionary paramGrid with arrays of criteria, maxDepth, minSamplesplit parameters
Output: dictionary: {criterion: best parameter in array, maxDepth: best parameter in array, minSamplesplit: best parameter in array}
1. define a tree object tree
2. flag←1, n←1
3. while flag==1 do
    4. flag←0
    5. define the grid search model GridSearchCV, passing in the dictionary paramGrid, the object tree, the evaluation metric roc_auc and the number of cross-validations 10;
    6. train the grid search model and use the grid search model for prediction and validation;
    7.p=optimal dictionary parameters for grid search under the current paramGrid: {criterion: optimal parameter i in the array, maxDepth: optimal parameter j in the array, minSamplesplit: optimal parameter m in the array};
    8. bringing p into the tree model for training
    9. auc2←tree results of the model evaluation
    10. if n==1 then
        11. auc1←tree results of the model evaluation
    12. else if auc2>auc1 then
        13. flag←1, auc1←auc2
    14. else
        15. output p
    16. end if
    17. update paramGrid←{criterion: ['entropy', 'gini'], maxDepth: [j-2, j-1, j, j+1, j+2], minSamplesplit: [m-2, m-1, m, m+1, m+2]};
    18. n++
19. end while
```

E. Modeling method

1) KNN: The KNN algorithm generally uses the majority voting method; that is, the majority classes of K neighbors of the input instance, to determine the class of the input instance [28]. To obtain the closest training data in the test data, the test range interval was calculated for each training data, which can be Euclidean, Mahalobins distance, Cosine, City block, Chebychev, Correlation (corr), Hamming, Jaccard, Minkowski, Seuclidean and Spearman [29], [30].

Selection of the K value: The K value determines the complexity and generalization ability of the model. The higher the K value, the lower the model complexity, the stronger the generalization ability, and the greater the training error [30]. The K value is generally <20, and this study adopted the cross verification method [31] to obtain the appropriate K value.

2) DT: The structure of the DT is shown in Fig. 1. The DT is divided into several branch nodes (classification attributes of the bifurcation path table) by the root node, and the data are classified into leaf nodes according to the size of the attribute values to complete the decision classification [32]. DT is used for multi-label classification, which assigns a series of labels to a specific sample and trains a classifier for each label. An extended estimator is then generated to evaluate a series of objective functions, which are trained on a separate prediction matrix to predict a series of response [19]. The model uses 10-fold cross-validation and grid search to optimize the model. k-fold cross-validation was used to divide the training samples into k parts, with one part as the data for model validation and the remaining k-1 part for training [31], [32]. Finally, all samples were verified once, and the k times results were averaged into a single estimated value. This improved the generalizability of the model and prevented overfitting. The grid search presets several parameter combinations, using cross validation to evaluate each group of parameters, select the best parameters to establish the model, and achieve the pruning effect [33].

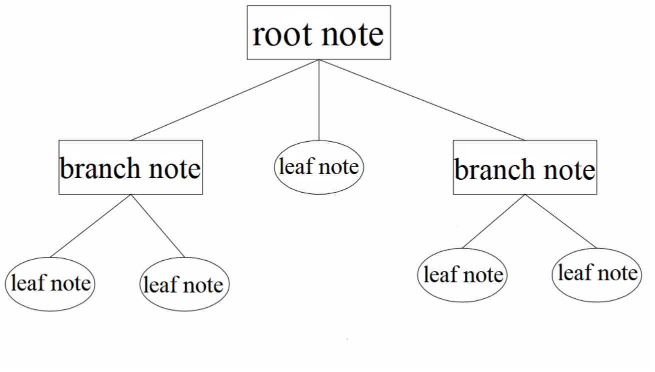

Fig. 1. DT structure

3) RF: The RF algorithm uses the bootstrap method to randomly sample N new self-help sample sets and create N regression trees, and is a robust classifier for the training and samples prediction, using multiple DTs. As shown in Fig. 2, each DT is used to make a judgment according to its own state and to vote, to select the final classification result, which overcomes the shortcoming of easy overfitting of a single-family DT [34]. The model also uses 10-fold cross-validation and grid search to adjust for the over-fitted parameters for multi-label classification.

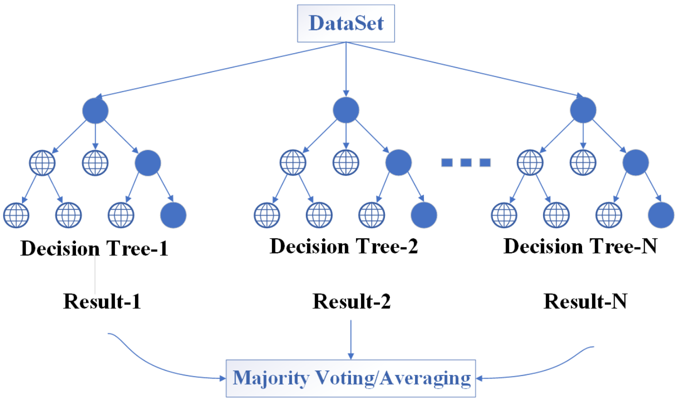

Fig. 2. RF structure

F. Model evaluation indicators

In machine learning, the accuracy, precision, recall, F1 score, area under the curve (AUC) and $weighted{-}F_1$ are commonly used to measure the correct classification ability of two-class classifiers. The values of the first four indicators ranged between [0 and 1]. The larger the value, the better the model effect. The formula used is as follows [35], [36], [37], [38]:

$$accuracy = \frac{TP + TN}{(TP + FN + FP + TN)} \tag{3}$$

$$precision = \frac{TP}{(TP + FP)} \tag{4}$$

$$recall = \frac{TP}{(TP + FN)} \tag{5}$$

$$F_1 = \frac{2(precision \cdot recall)}{(precision + recall)} \tag{6}$$

$$weighted{-}F_1 = \sum_{i=1}^{n} F_i w_i \tag{7}$$

where TP, TN, FP, and FN denotes the number of correctly predicted positive and negative, and incorrectly predicted positive and negative cases, respectively. The F1 score, which considers both the accuracy and recall rates of the classification model, can be regarded as a weighted average of the model accuracy and recall rates [37]. The AUC value [between 0.5, 1] represents the probability of judging a sample as having a positive score, and is greater than a negative sample score [37]. The larger the value, the better the model prediction performance. The $weighted{-}F_1$ takes into account the number of samples in each category as a proportion of the total sample, and can ignore data imbalances to some extent [39].

III. RESULTS

A. Model prediction process

In this study, three machine learning algorithms were used to analyze primary BC data obtained from the SEER database. After data selection and transformation, oversampling and balancing, cyclic parameter tuning, model construction, and multi-label classification, the prediction process were completed as shown in Fig. 3.

TABLE IV
DIRECTION OF METASTASIS

Metastatic site	The total sample size	Total proportion (%)	Percentage of metastasis direction (%)
Not metastasis (metsTotal=0)	282,369	96.0615	/
Bone metastasis (unidirectional)	7,456	2.5365	34.6629
Lung metastasis (unidirectional)	3,367	1.1454	15.6532
Liver metastasis (unidirectional)	2,871	0.9767	13.3473
Bone + lung	1,841	0.6263	8.5588
Bone + liver	1,647	0.5603	7.6569
Lung + liver	964	0.3280	4.4816
Brain metastases (unidirectional)	808	0.2749	3.7564
Bone + lung + liver	671	0.2283	3.1195
Bone + brain	511	0.1738	2.3756
Lung + brain	373	0.1269	1.7341
Liver + brain	261	0.0888	1.2134
Bone + lung + brain	248	0.0844	1.1530
Bone + liver + brain	199	0.0678	0.9252
Lung + liver + brain	162	0.0551	0.7531
Bone + lung + liver + brain	131	0.0446	0.6090

B. Direction of metastasis

The original data were analyzed to detect metastasis before model construction, and the direction of metastasis was obtained, as shown in Table IV.

C. Construction of the prediction model

After mice package was interpolated several times, MLSMOTE was used to oversample the balanced sample data. The train_test_split function was used to create a 7/3 (70% as a training set and 30% as a verification set) data-balancing split [35]. Twenty-one BC attribute fields in the data collected were used as model predictive variables, while MetsTotal, MetsBone, MetsLung, MetsLiver, MetsBrain attributes were used as binary multi-label result variables. A variety of machine learning methods (KNN, DT, and RF) were used to construct the model, and the performance of the classifier in the validation set was evaluated based on the accuracy, precision, recall, F1 score, AUC and $weighted{-}F_1$ indicators. Finally, the optimal prediction model and main attributes affecting the direction of BC metastasis were selected and derived.

When patients with primary BC participate in treatment, physicians can choose the prediction model with the best evaluation effect, input the specific attribute field values of patients (influencing factors of metastasis direction) in the order of importance; and obtain the output results to predict the signs of cancer metastasis in patients. If the model predicts MetsTotal=1 (metastasis), physicians can perform modified radical mastectomy; and the possible direction of metastasis could be monitored during follow-up (whether MetsBone, MetsLung, MetsLiver, MetsBrain are 1, in which case, close follow-up should be conducted for cancers at specific sites). If MetsTotal=0 (no metastasis), axillary dissection with periodic screening is recommended as the primary

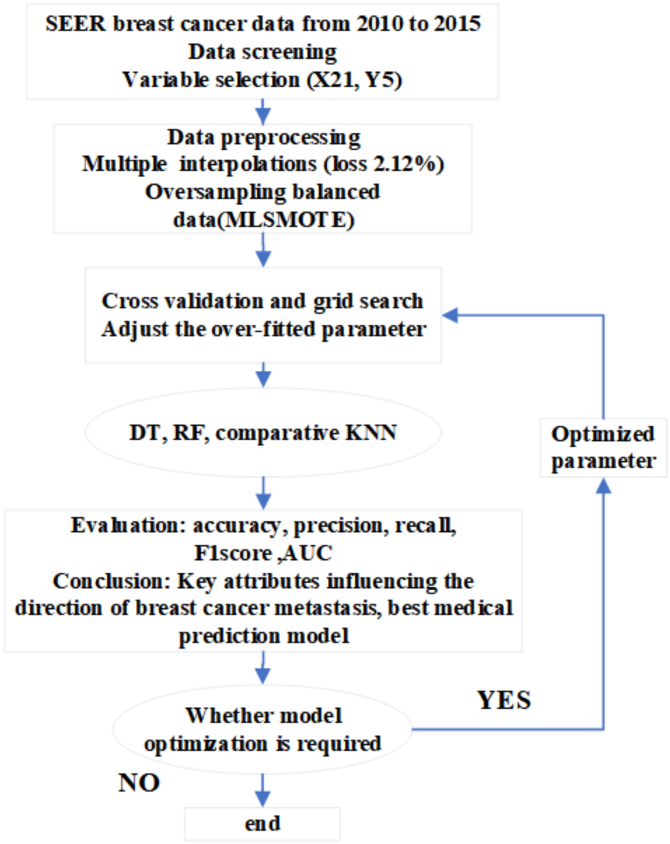
SEER breast cancer data from 2010 to 2015
Data screening
Variable selection (X21, Y5)
Data preprocessing
Multiple interpolations (loss 2.12%)
Oversampling balanced data(MLSMOTE)
Cross validation and grid search
Adjust the over-fitted parameter
DT, RF, comparative KNN
Optimized parameter
Evaluation: accuracy, precision, recall, F1score ,AUC
Conclusion: Key attributes influencing the direction of breast cancer metastasis, best medical prediction model
Whether model optimization is required
YES
NO
end

Fig. 3. The modeling process

follow-up strategy. The actual surgical treatment and follow-up plan were based on the actual condition of the patients and the experience of the physicians, and the results of the study only serve as an auxiliary reference.

D. Parameter tuning results under improved fully automated cross grid

The improved fully automated cross grid algorithm adjusted the size of the super-parameter based on the model indicators (accuracy, precision, recall, F1 score, AUC and $weighted{-}F_1$). After repeated circulation adjustment, the maximum depth range of the DT was selected [6-11], the index range of the split dataset [entropy, gini], and the minimum number of split leaf samples [4, 8, 12, 16, 20, 24] were indicated. An exhaustive search was used to select the optimal parameters for (max_depth: 9; criterion: entropy; min_samples_split: 16).

Similarly, the RF also underwent multiple tuning. Finally, the criterion (split data set metrics) was entropy, max_depth (the maximum depth of each DT) was 9, min_samples_split (minimum split sample size of leaves per tree) was 18, and n_estimators (number of DTs) was 11.

E. Analysis of model results

In this study, DT and RF were used to train real-world datasets, and the importance scores of the attributes affecting the direction of metastasis were obtained, as shown in Table V. Surgical information (SurgPrim) and lymph node status (RNP, RNE) had a greater influence on organ metastasis in primary BC. The existence of distant metastasis (AJCCM), tumor size (TumorSize), chemotherapy (Chemotherapy), histologic type (Histologic), radiation (Radiation), and age at diagnosis (Age) showed good characteristic expressions for the probability of organ metastasis in primary BC. These indicate a good reference for the organ metastasis study in patients with primary BC.

The prediction results of KNN, DT, and RF are listed in Tables VI and VII. For evaluation indexes, the highest score of the RF of the six labels occurred in the accuracy index, with an average score of 94.14%, higher than that of DT (average score of 93.59%) and KNN (average score of 88.84%). Among the precision index evaluation results, DT scored the highest except with the lung metastasis label, with an average score of 95.85%, which was 1.11% and 8.61% higher than for RF and KNN, respectively. In the recall index evaluation results, DT had the best effect in predicting total, lung, and brain metastases; RF had the highest score except for lung metastasis; while the RF model had the highest average score (91.76%). Among the F1 score evaluation results, DT had the best effect in predicting total, bone, and brain metastases; RF had the highest score except for bone metastasis; while the RF model had the highest average score (93.45%). In the AUC index evaluation results, RF scored the highest, except for the total metastasis label, with an average score of 99.95%, which was 0.42% and 2.84% higher compared to those of DT and KNN, respectively. RF's $weighted{-}F_1$ metrics increased by 1.04% and 8.35% compared to DT and KNN, respectively. Generally, RF performed well in the evaluation of the six indicators, followed by DT and KNN. This is because KNN, based on the regression analysis, cannot process highly correlated and non-linear data, and the correlation between BC data variables was high; therefore, the prediction effect of the model was not ideal. Although the DT algorithm could deal with the highly correlated data and the prediction result was significantly improved, a single tree is easy to fit; therefore, the strong classifier RF algorithm composed of multiple DTs, showed the best effect.

Receiver Operating Characteristic (ROC) prediction curves of bone, lung, liver, and brain metastases of the three machine learning algorithms are shown in Figs. 4. Among the four y tag predictions, the RF curves were the closest to the upper-left corner, showing the best performance.

IV. DISCUSSION

In this study, multi-label prediction of the metastatic direction of primary BC patients from 2010 to 2015 based on the SEER database was examined. Three machine learning methods (KNN, DT, and RF) were used to construct the models. The evaluation index value of the model was high; this can effectively assist the doctors in early diagnosis and treatment. MLSMOTE and machine learning were used to balance the data and present the model data relationship. Class imbalance is a common actual classification problem, and most machine learning algorithms undertake data balance as the premise of practice. Therefore, class imbalance can introduce great challenges to the prediction task. The oversampling method can effectively overcome the problem of data imbalance by overemphasizing the positive proportional data (repeating the positive proportional data but not introducing more data into the model) and by enhancing the impact of the

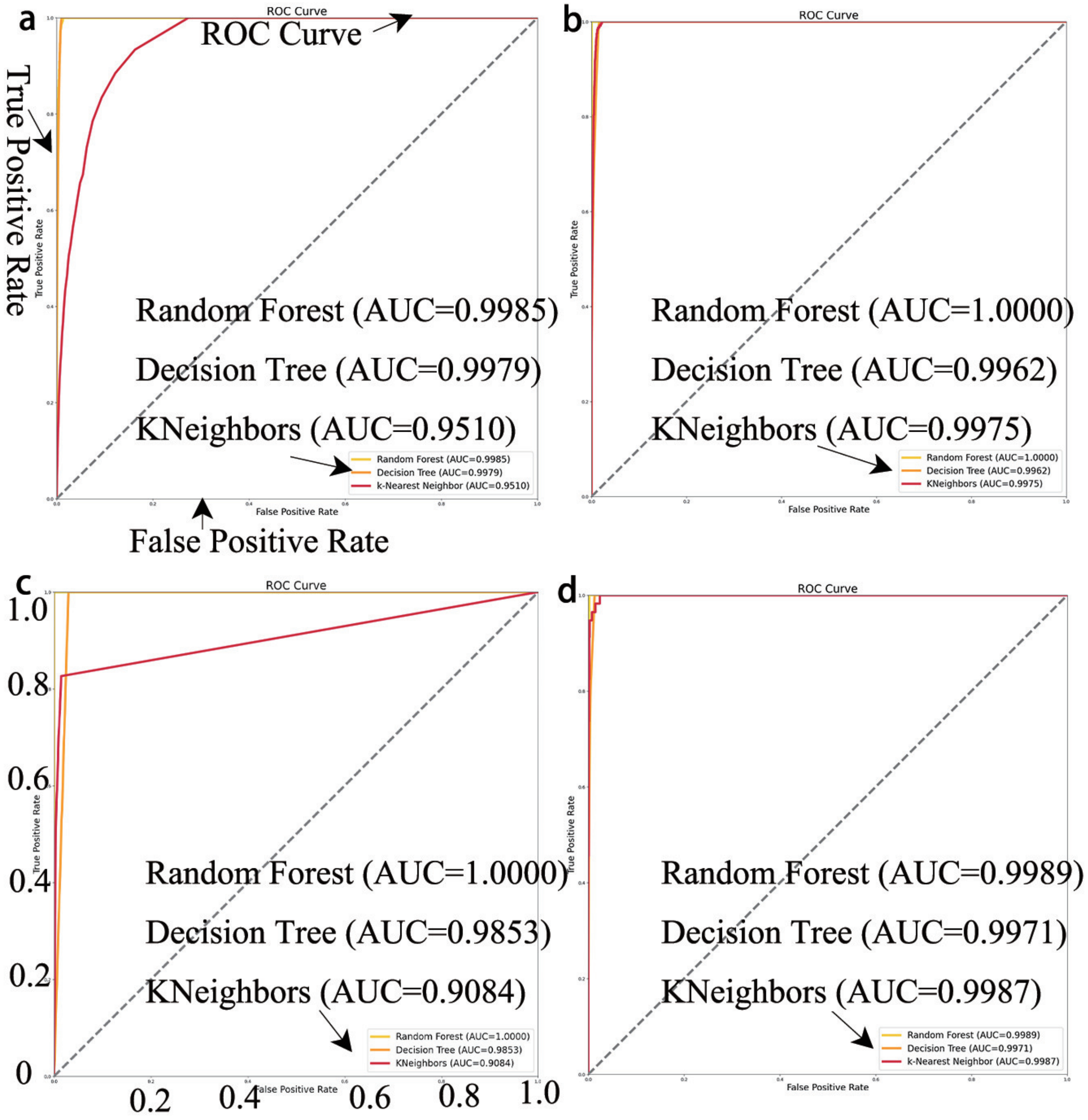

Fig. 4. ROC curve of BC metastases. Panel a shows ROC curve of bone metastasis. Panel b shows ROC curve of lung metastasis. Panel c shows ROC curve of liver metastasis. Panel d shows ROC curve of brain metastasis.

positive proportional noise on the model [37]. The study preprocessed data MetsTotal label (1:0=3.94:96.06), MetsBone label (1:0=2.54:97.46), MetsLung label (1:0=1.15:98.85), MetsLiver label (1:0=0.98:99.02), and MetsBrain label (1:0=0.27:98.75) were all out of proportion. However, the common oversampling method, such as the SMOTE, cannot deal with the multi-label problems; thus the need to introduce the optimized MLSMOTE, to process the imbalanced data [26], [27]. Therefore, in characteristic engineering, use of the MLSMOTE to balance multi-label problem data, can provide a reference for machine learning engineers.

The model evaluation in this study was based on accuracy, precision, recall, F1 score, AUC, and $weighted{-}F_1$, to measure the multi-label classification ability of the machine learning methods. The average accuracy of the RF and DT models was 93.59% and 94.14%, average precision was 95.85% and 94.74%, average recall was 90.40% and 91.76%, average F1 score was 92.16% and 93.45%, average AUC was 99.53% and 99.95%, $weighted{-}F_1$ was 89.44% and 99.48%, respectively. The prediction effect was better than that of deep learning or image processing methods, which have become popular in recent years. Zheng et al. reported deep learning radiomics of conventional ultrasound and shear wave elastography of BC, to predict axillary lymph node (ALN) status preoperatively in patients with early stage BC [40]. With an AUC of 90.50%, this was lower than the 99.95% for RF reported in this study. Liu et al. extracted quantitative imaging features from T2-weighted, diffusion-weighted, and contrast-enhanced T1-weighted images prior to each patient's NAC, to predict the pathological complete

TABLE V
ATTRIBUTE SCORE TABLE

Variable	Decision tree (%)	Random forest (%)
Race	0.1472	0.4251
Age	**0.2314**	**1.0150**
YD	<0.001	0.2530
Marital	<0.001	0.3773
Laterality	<0.001	0.4259
PrimarySite	<0.001	0.6785
Chemotherapy	**0.1760**	**4.2529**
Radiation	**1.7076**	**1.2061**
Grade	<0.001	0.5128
Histologic	**0.4685**	**2.2678**
LymphNodes	0.2574	0.9863
RNE	**0.4382**	**11.5141**
RNP	**0.3337**	**3.9956**
AJCCT	0.1810	0.5863
AJCCN	0.1682	0.4259
AJCCM	**0.4790**	**5.5254**
HER2	<0.001	0.7984
ER	<0.001	0.6518
PR	<0.001	0.7374
SurgPrim	**95.0756**	**58.8750**
TumorSize	**0.3357**	**4.4688**

TABLE VI
MACHINE LEARNING CLASSIFICATION PERFORMANCE COMPARISON

Multi-label y	**Accuracy (%)**			**Precision (%)**		
	KNN	**DT**	**RF**	**KNN**	**DT**	**RF**
Class 0 (MetsTotal)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Class 1 (MetsBone)	79.56	**90.61**	**90.61**	81.02	**97.14**	90.98
Class 2 (MetsLung)	79.56	86.74	**87.29**	78.69	82.11	**87.62**
Class 3 (MetsLiver)	85.08	90.61	**92.82**	76.47	**100.00**	95.12
Class 4 (MetsBrain)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Ave	88.84	93.59	**94.14**	87.24	**95.85**	94.74

response preconditioning of NAC for BC, based on the proposed multi-parameter magnetic resonance imaging radiometrics [41]. The AUC assessment result of 86% was lower than the 99.95% for RF model in this study. Muhammad et al. created a multilayer feed-forward neural network model based on proportional conjugate gradient backpropagation; and used salivary amino acid bioaccurate labelling of gastric cancer with an average accuracy of 92.27%, but still lower than the 94.14% in this study [42].

Early studies have focused on the gene detection technology innovation on BC metastasis studies. To the best of our knowledge, this is the first machine-learning-based approach to predict the direction of organ metastasis in BC. By constructing a multi-label classification model, this study findings can be used to identify the relationship between characteristic attributes and prognosis of organ metastasis and assist doctors in diagnosis and treatment, which has theoretical and medical significance. Currently, gene detection technology is often used in clinical medicine to identify mutated genes, assist in monitoring disease metastasis, and guide treatment selection [43]. Cancer cell samples obtained during genetic testing are mostly from surgical pathological sections, and biopsy and ”liquid biopsies”. The method of obtaining surgical pathological sections is unsuitable for preoperative detection or long-term postoperative follow-up (2 years later) [10]. Puncture biopsy is highly invasive and risky, and ”liquid biopsy” is most suitable for patients with advanced cancer [10], [11]. Kim et al. reported an unusual case of a patient that was pathologically described as a primary hypercellular parathyroid lesion with characteristic changes on fine-needle aspiration (FNA) biopsy [44]. These results suggest that FNA can enhance the similarity between malignant tumors, with challenging diagnoses. Therefore, there is an urgent need for new tests to replace such invasive and non-universal tests. Arefan et al. constructed a machine learning classifier to distinguish positive and negative ALN status of BC and achieved high classification performance, showing convenience, non-invasive, universal, and repeatable advantages [45]. Therefore, since the analysis of routine clinicopathological features of BC patients using a machine learning algorithm is non-invasive and universal, it can assist in the assessment of patients level of severity and provide a reference for clinical practice.

Although our study performed well in identifying the direction of BC metastasis, it has some limitations. (1) Among independent variables, Ki-67, patients’ psychologic, family, and other factors are important and affect the direction of BC metastasis; however, the SEER database does not contain information on these important factors [46]. With dependent variables, the attribute value of the lymphatic metastasis variable was missing, resulting in only six types of y values [47]. (2) Dong et al. found that BC and thyroid cancer (TC) tend to occur heterochronously or synchronously [48]. Because both glands are regulated by the hypothalamic-pituitary axis, the correlation between the two is an important consideration in cancer research. However, in the prediction of cancer cell metastasis direction in this study, the potential correlation between BC and other cancers was not considered, such as whether the probability of cancer cell metastasis would be higher in simultaneous first-stage BC patients with TC. This requires the collection of relevant data on cancer types by a physician, using their commonalities and differences in prognostic factors to assess the association. At present, the authors showed that the prognostic factors of BC and TC are highly similar and may be associated. In future, a controlled experiment will be conducted based on the data of patients with first-episode BC alone and TC combined with first-episode BC, to analyze the direction and influencing factors of cancer cell metastasis and explore whether TC has an impact on the prognosis of first-episode BC metastasis. (3) In this study, only the internal validation method of cross validation was adopted. Although the repeatability of the model was effectively verified, its generalization and portability are yet unverified. It is necessary to simultaneously adopt both internal and external verifications. Based on the good performance of the internal verification of the development model, external verifications such as spatial, domain, and period verifications should be performed.

This study, based on the first primary BC diagnosis and treatment of patients, was conducted to predict the direction of future cancer metastasis, and to assist in the formulation of treatment and follow-up plans. However, in actual medical scenarios, when BC patients are diagnosed and undergo treatment, metastasis may occur, and BC may not be the primary

TABLE VII
MACHINE LEARNING CLASSIFICATION PERFORMANCE COMPARISON

Multi-label y	Recall (%)			F1 Score (%)			AUC (%)		
	KNN	DT	RF	KNN	DT	RF	KNN	DT	RF
Class 0 (MetsTotal)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Class 1 (MetsBone)	90.98	87.93	**97.39**	85.71	**92.31**	91.29	95.10	99.79	**99.85**
Class 2 (MetsLung)	89.71	**98.06**	90.00	83.84	88.99	**91.87**	99.75	99.62	**100.00**
Class 3 (MetsLiver)	57.78	66.00	**71.43**	65.82	79.52	**84.09**	90.84	98.53	**100.00**
Class 4 (MetsBrain)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.87	99.71	**99.89**
Ave	87.73	90.40	**91.76**	87.07	92.16	**93.45**	97.11	99.53	**99.95**

TABLE VIII
MACHINE LEARNING CLASSIFICATION PERFORMANCE COMPARISON

Evaluation	KNN	DT	RF
$weighted{-}F_1$	82.13%	89.44%	**90.48%**

cancer (i.e., when the primary site cannot be determined despite standard diagnostic tests) [49]. The knowledge of a patient's primary cancer site is fundamental to medical treatment; therefore, patients with unknown primary BC site are significantly disadvantaged, and with poor survival outcomes for most [50]. The development of reliable and easily accessible diagnostic methods to predict the origin of cancer tissues has become an important research topic in the medical field. Zhao et al. developed an RNA-based classifier called CUP-AI-DX that utilizes a 1D inception convolutional neural network (1D-inception) model to infer the primary tissue of the tumor [51]. The overall accuracy, based on The Cancer Genome Atlas Project and International Cancer Genome Consortium was 98.54% and 96.70%, respectively. Therefore, our next research aims to use machine learning methods to build a model to infer the origin of cancer tissues in patients with unknown primary BC site, and fully consider the potential correlation between BC and other cancers.

REFERENCES

[1] Y. Liang, H. Zhang, X. Song, and Q. Yang, "Metastatic heterogeneity of breast cancer: Molecular mechanism and potential therapeutic targets." *Semin Cancer Biol*, vol. 60, no. 1, pp. 14–27, 2020.

[2] O. Pagani, E. Senkus, W. Wood, M. Colleoni, T. Cufer, S. Kyriakides, A. Costa, E. Winer, and F. Cardoso, "International guidelines for management of metastatic breast cancer: can metastatic breast cancer be cured?" *J Natl Cancer Inst*, vol. 102, no. 7, pp. 456–63, 2010.

[3] C. G. A. Network, "Comprehensive molecular portraits of human breast tumours." *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[4] R. Wang, Y. Zhu, X. Liu, X. Liao, J. He, and L. Niu, "The clinicopathological features and survival outcomes of patients with different metastatic sites in stage iv breast cancer." *BMC Cancer*, vol. 19, no. 1, pp. 1091–1108, 2019.

[5] C. Allemani, T. Matsuda, and C. V. Di, "Global surveillance of trends in cancer survival 2000-14 (concord-3): analysis of individual records for 37513025 patients diagnosed with one of 18 cancers from 322 population based registries in 71 countries." *Lancet*, vol. 391, no. 10125, pp. 1023–1075, 2018.

[6] S. Valastyan and R. Weinberg, "Tumor metastasis: molecular insights and evolving paradigms." *Cell*, vol. 147, no. 2, pp. 275–92, 11 2011.

[7] M. Puppo, M. Valluru, and P. Clézardin, "Micrornas and their roles in breast cancer bone metastasis." *Curr Osteoporos Rep*, vol. 19, no. 3, pp. 256–263, 2021.

[8] T. N. Satomi, A. Shimomura, J. Matsuzaki, Y. Yamamoto, J. Kawauchi, S. Takizawa, Y. Aoki, H. Sakamoto, K. Kato, C. Shimizu, T. Ochiya, and K. Tamura, "Serum microrna-based prediction of responsiveness to eribulin in metastatic breast cancer." *PLoS One*, vol. 14, no. 9, pp. 222 024–222 035, 2019.

[9] X. Feng, M. Zhang, B. Wang, C. Zhou, Y. Mu, J. Li, X. Liu, Y. Wang, Z. Song, and P. Liu, "Crabp2 regulates invasion and metastasis of breast cancer through hippo pathway dependent on er status." *J Exp Clin Cancer Res*, vol. 38, no. 1, pp. 361–378, 2019.

[10] J. Lidbury, "Getting the most out of liver biopsy." *Vet Clin North Am Small Anim Pract*, vol. 47, no. 3, pp. 569–583, 2017.

[11] P. Pisapia, U. Malapelle, and G. Troncone, "Liquid biopsy and lung cancer," *Acta Cytol*, vol. 63, no. 6, pp. 489–496, 2019.

[12] J. Holm, J. Li, H. Darabi, M. Eklund, M. Eriksson, K. Humphreys, P. Hall, and K. Czene, "Associations of breast cancer risk prediction tools with tumor characteristics and metastasis." *J Clin Oncol*, vol. 34, no. 3, pp. 251–258, 2016.

[13] X. Cheng, L. Xia, and S. Sun, "A pre-operative mri-based brain metastasis risk-prediction model for triple-negative breast cancer." *Gland Surg*, vol. 10, no. 9, pp. 2715–2723, 2021.

[14] A. Yang, W. Xiao, S. Zheng, Y. Kong, Y. Zou, and M. Li, "Predictive nomogram of subsequent liver metastasis after mastectomy or breast-conserving surgery in patients with nonmetastatic breast cancer." *Cancer Control*, vol. 28, no. 1, pp. 1–7, 2021.

[15] A. Giuliano, S. Edge, and G. Hortobagyi, "Eighth edition of the ajcc cancer staging manual: Breast cancer." *Ann Surg Oncol*, vol. 25, no. 7, pp. 1783–1785, 2018.

[16] D. Teichgraeber, M. Guirguis, and G. Whitman, "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*, vol. 217, no. 2, p. 278–290, 2021.

[17] C. Mercan, S. Aksoy, E. Mercan, L. Shapiro, D. Weaver, and J. Elmore, "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images." *IEEE Trans Med Imaging*, vol. 37, no. 1, pp. 316–325, 2018.

[18] Y. Qu, G. Yue, C. Shang, L. Yang, R. Zwiggelaar, and Q. Shen, "Multi-criterion mammographic risk analysis supported with multi-label fuzzy-rough feature selection." *Artif Intell Med*, vol. 100, no. 1, pp. 101 722–101 735, 2019.

[19] E. Tanaka, S. Nozawa, A. Macedo, and J. Baranauskas, "A multi-label approach using binary relevance and decision trees applied to functional genomics." *J Biomed Inform*, vol. 54, no. 1, pp. 85–95, 2015.

[20] F. SJ and T. A, "Exploiting medline for gene molecular function prediction via nmf based multi-label classification." *J Biomed Inform*, vol. 86, no. 1, pp. 160–166, 2018.

[21] L. Zhou, X. Zheng, D. Yang, Y. Wang, X. Bai, and X. Ye, "Application of multi-label classification models for the diagnosis of diabetic complications." *BMC Med Inform Decis Mak*, vol. 21, no. 1, pp. 182–191, 2021.

[22] M. Daly and I. Paquette, "Surveillance, epidemiology, and end results (seer) and seer-medicare databases: Use in clinical research for improving colorectal cancer outcomes." *Clin Colon Rectal Surg*, vol. 32, no. 1, pp. 61–68, 2019.

[23] A. Pedersen, E. Mikkelsen, F. D. Cronin, N. Kristensen, T. Pham, and I. Pedersen, Land Petersen, "Missing data and multiple imputation in clinical epidemiological research." *Clinical Epidemiology*, vol. 9, no. 1, pp. 157–166, 2017.

[24] B. Leurent, M. Gomes, S. Cro, N. Wiles, and J. Carpenter, "Reference-based multiple imputation for missing data sensitivity analyses in trial-based cost-effectiveness analysis." *Health Econ*, vol. 29, no. 2, pp. 171–184, 2020.

[25] G. Cserni, E. Chmielik, B. Cserni, and T. Tot, "The new tnm-based staging of breast cancer." *Virchows Arch*, vol. 472, no. 5, pp. 697–703, 2018.

[26] W. Lin and D. Xu, "Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types." *Bioinformatics*, vol. 32, no. 24, pp. 3745–3752, 2016.

[27] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on smote algorithm." *Sci Rep*, vol. 11, no. 1, pp. 24 039–24 049, 2021.

[28] T. Wang, Q. Fan, H. Cai, and B. Zhang, “Application of machine learning for tracing the origin of metastatic lung cancer tissues.” *IAENG International Journal of Computer Science*, vol. 50, no. 2, pp. 359–367, 2023.

[29] L. Zakaria, H. M. Ebeid, and S. Dahshan, “Analysis of classification methods for gene expression data.” *Advances in lntelligent Systems and Computing*, vol. 921, no. 1, pp. 190–199, 2020.

[30] M. Tahir, M. Hayat, and M. Kabir, “Sequence based predictor for discrimination of enhancer and their types by applying general form of chou’s trinucleotide composition.” *Comput Methods Programs Biomed*, vol. 146, no. 1, pp. 69–75, 2017.

[31] M. Montazeri and A. Beigzadeh, “Machine learning models in breast cancer survival prediction.” *Technol Health Care*, vol. 24, no. 1, pp. 31–42, 2016.

[32] C. Chao, Y. Yu, B. Cheng, and Y. Kuo, “Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree.” *J Med Syst*, vol. 38, no. 10, p. 106, 2014.

[33] H. Zhou, Z. Zhong, M. Hu, and J. Huang, “Determining the steering direction in critical situations: A decision tree-based method.” *Traffic Inj Prev*, vol. 21, no. 6, pp. 395–400, 2020.

[34] B. Macaulay, B. Aribisala, S. Akande, B. Akinnuwesi, and O. Olabanjo, “Breast cancer risk prediction in african women using random forest classifier.” *Cancer Treat Res Commun*, vol. 28, no. 1, pp. 100 396–100 402, 2021.

[35] Y. Xu, L. Jiao, S. Wang, J. Wei, Y. Fan, M. Lai, and E. Chang, “Multi-label classification for colon cancer using histopathological images.” *Microsc Res Tech*, vol. 76, no. 12, pp. 1266–77, 2013.

[36] R. Arian, A. Hariri, A. Mehridehnavi, A. Fassihi, and F. Ghasemi, “Protein kinase inhibitors’ classification using k-nearest neighbor algorithm.” *Comput Biol Chem*, vol. 86, no. 1, pp. 107 269–107 279, 2020.

[37] Z. DeVries, E. Locke, M. Hoda, D. Moravek, K. Phan, A. Stratton, S. Kingwell, E. Wai, and P. Phan, “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*, vol. 21, no. 7, pp. 1135–1142, 2021.

[38] M. Asif, M. M. Nishat, F. Faisal, R. R. Dip, M. H. Udoy, M. Shikder, R. Ahsan *et al.*, “Performance evaluation and comparative analysis of different machine learning algorithms in predicting cardiovascular disease.” *Engineering Letters*, vol. 29, no. 2, pp. 731–741, 2021.

[39] G. Vandewiele, I. Dehaene, G. Kovács, L. Sterckx, O. Janssens, F. Ongenae, B. F. De, T. F. De, K. Roelens, J. Decruyenaere, H. S. Van, and T. Demeester, “Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling.” *Artif Intell Med*, vol. 111, no. 1, pp. 101 987–101 997, 2021.

[40] X. Zheng, Z. Yao, Y. Huang, Y. Yu, Y. Wang, Y. Liu, R. Mao, F. Li, Y. Xiao, Y. Wang, Y. Hu, J. Yu, and J. Zhou, “Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer.” *International Journal of Molecular Sciences*, vol. 11, no. 1, pp. 1236–1244, 2020.

[41] Z. Liu, Z. Li, J. Qu, R. Zhang, X. Zhou, L. Li, K. Sun, Z. Tang, H. Jiang, H. Li, Q. Xiong, Y. Ding, X. Zhao, K. Wang, Z. Liu, and J. Tian, “Radiomics of multiparametric mri for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: A multicenter study.” *Clin Cancer Res*, vol. 25, no. 12, pp. 3538–3547, 2019.

[42] M. A. Aslam, C. Xue, M. Liu, K. Wang, and D. Cui, “Classification and prediction of gastric cancer from saliva diagnosis using artificial neural network.” *Engineering Letters*, vol. 29, no. 1, pp. 10–24, 2020.

[43] M. Dameri, L. Ferrando, G. Cirmena, C. Vernieri, G. Pruneri, A. Ballestrero, and G. Zoppoli, “Multi-gene testing overview with a clinical perspective in metastatic triple-negative breast cancer.” *Int J Mol Sci*, vol. 22, no. 13, pp. 7154–7177, 2021.

[44] J. Kim, G. Horowitz, M. Hong, M. Orsini, S. Asa, and K. Higgins, “The dangers of parathyroid biopsy.” *J Otolaryngol Head Neck Surg*, vol. 46, no. 1, pp. 4–7, 2017.

[45] D. Arefan, R. Chai, M. Sun, M. Zuley, and S. Wu, “Machine learning prediction of axillary lymph node metastasis in breast cancer: 2d versus 3d radiomic features.” *Int J Mol Sci*, vol. 47, no. 12, pp. 6334–6342, 2020.

[46] Y. Yin, K. Zeng, M. Wu, Y. Ding, M. Zhao, and Q. Chen, “The levels of ki-67 positive are positively associated with lymph node metastasis in invasive ductal breast cancer.” *Cell Biochem Biophys*, vol. 70, no. 2, pp. 1145–51, 2014.

[47] B. To, D. Isaac, and E. Andrechek, “Studying lymphatic metastasis in breast cancer: Current models, strategies, and clinical perspectives.” *J Mammary Gland Biol Neoplasia*, vol. 25, no. 3, pp. 191–203, 2020.

[48] L. Dong, J. Lu, B. Zhao, W. Wang, and Y. Zhao, “Review of the possible association between thyroid and breast carcinoma.” *World J Surg Oncol*, vol. 16, no. 1, pp. 130–136, 2018.

[49] X. Liu, L. Li, L. Peng, B. Wang, J. Lang, Q. Lu, X. Zhang, Y. Sun, G. Tian, H. Zhang, and L. Zhou, “Predicting cancer tissue-of-origin by a machine learning method using dna somatic mutation data.” *Front Genet*, vol. 11, no. 1, pp. 674–684, 2020.

[50] D. Lu, J. Jiang, X. Liu, H. Wang, S. Feng, X. Shi, Z. Wang, Z. Chen, X. Yan, H. Wu, and K. Cai, “Machine learning models to predict primary sites of metastatic cervical carcinoma from unknown primary.” *Front Genet*, vol. 11, no. 1, pp. 614 823–614 830, 2020.

[51] Y. Zhao, Z. Pan, S. Namburi, A. Pattison, A. Posner, S. Balachander, C. Paisie, H. Reddi, J. Rueter, A. Gill, S. Fox, K. Raghav, W. Flynn, R. Tothill, S. Li, R. Karuturi, and J. George, “Cup-ai-dx: A tool for inferring cancer tissue of origin and molecular subtype using rna gene-expression data and artificial intelligence.” *EBioMedicine*, vol. 61, no. 1, pp. 103 030–103 043, 2020.