frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 07 March 2023
DOI 10.3389/fonc.2023.1117420

OPEN ACCESS

EDITED BY
Rajesh Kumar Tripathy,
Birla Institute of Technology and Science, India

REVIEWED BY
Pranjali Gajbhiye,
Nirvesh Enterprises Private Limited, India
Jun Jiang,
Mayo Clinic, United States

*CORRESPONDENCE
Weiwei Yang
yangweiwei113@163.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Breast Cancer, a section of the journal Frontiers in Oncology

RECEIVED 08 December 2022
ACCEPTED 16 February 2023
PUBLISHED 07 March 2023

CITATION
Zeng L, Liu L, Chen D, Lu H, Xue Y, Bi H and Yang W (2023) The innovative model based on artificial intelligence algorithms to predict recurrence risk of patients with postoperative breast cancer.
Front. Oncol. 13:1117420.
doi: 10.3389/fonc.2023.1117420

The innovative model based on artificial intelligence algorithms to predict recurrence risk of patients with postoperative breast cancer

Lixuan Zeng[1†], Lei Liu[2†], Dongxin Chen[1†], Henghui Lu[3†], Yang Xue[1], Hongjie Bi[1] and Weiwei Yang[1*]

[1]Department of Pathology, Harbin Medical University, Harbin, China, [2]Department of Breast Surgery, The Third Affiliated Hospital of Harbin Medical University, Harbin, China, [3]Department of Dermatology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China

Purpose: This study aimed to develop a machine learning model to retrospectively study and predict the recurrence risk of breast cancer patients after surgery by extracting the clinicopathological features of tumors from unstructured clinical electronic health record (EHR) data.

Methods: This retrospective cohort included 1,841 breast cancer patients who underwent surgical treatment. To extract the principal features associated with recurrence risk, the clinical notes and histopathology reports of patients were collected and feature engineering was used. Predictive models were next conducted based on this important information. All algorithms were implemented using Python software. The accuracy of prediction models was further verified in the test cohort. The area under the curve (AUC), precision, recall, and F1 score were adopted to evaluate the performance of each model.

Results: A training cohort with 1,289 patients and a test cohort with 552 patients were recruited. From 2011 to 2019, a total of 1,841 textual reports were included. For the prediction of recurrence risk, both LSTM, XGBoost, and SVM had favorable accuracies of 0.89, 0.86, and 0.78. The AUC values of the micro-average ROC curve corresponding to LSTM, XGBoost, and SVM were 0.98 $\pm$ 0.01, 0.97 $\pm$ 0.03, and 0.92 $\pm$ 0.06. Especially the LSTM model achieved superior execution than other models. The accuracy, F1 score, macro-avg F1 score (0.87), and weighted-avg F1 score (0.89) of the LSTM model produced higher values. All *P* values were statistically significant. Patients in the high-risk group predicted by our model performed more resistant to DNA damage and microtubule targeting drugs than those in the intermediate-risk group. The predicted low-risk patients were not statistically significant compared with intermediate- or high-risk patients due to the small sample size (188 low-risk patients were predicted *via* our model, and only two of them were administered chemotherapy alone after surgery). The prognosis of patients predicted by our model was consistent with the actual follow-up records.

Conclusions: The constructed model accurately predicted the recurrence risk of breast cancer patients from EHR data and certainly evaluated the chemoresistance and prognosis of patients. Therefore, our model can help clinicians to formulate the individualized management of breast cancer patients.

KEYWORDS

breast cancer, recurrence risk, LSTM, XGBoost, SVM

Introduction

According to estimates from the Global Cancer Observatory (GLOBOCAN) in 2020, the incidence of female breast cancer ranked first, surpassing even lung cancer (1). Meanwhile, in China, the incidence of breast cancer has risen to the fourth among all cancer types and shows a trend of younger age (2). Breast cancer seriously harms women's life and health. Accurately evaluating the recurrence risk of postoperative breast cancer patients can greatly improve their prognosis through appropriate treatment (3).

With the digitization of medical information, machine learning models have been applied in oncology (4–6). In 2021, artificial intelligence (AI) was used to predict the occurrence of breast cancer metastasis by learning from clinical electronic health record (EHR) data to support individualized diagnosis for patients (7). EHRs contain numerous longitudinal records, including histopathology, molecular markers related to breast cancer, radiology, and clinical information. However, the manual integration of prognostic information from EHRs by clinical experts is time-consuming, laborious, and costly (8, 9). Therefore, precisely assessing the recurrence risk and improving the efficiency of clinical evaluation plays a crucial role in controlling the disease burden of breast cancer.

Support vector machine (SVM) is a powerful learning algorithm that is capable of addressing various dimensions of data through different kernel functions. For example, breast cancer cells were classified *in vitro* with an accuracy of 93% using linear and radial basis function (RBF) kernel SVMs (10). Extreme gradient boosting (XGBoost) is a decision tree-based algorithm that is widely used in machine learning. It minimizes the loss function of the model through a gradient descent algorithm and implements the speed and performance of gradient-boosted decision trees (11). Furthermore, artificial neural networks (ANN) comprise a fundamental component of deep learning algorithms, demonstrating great potential in building high prediction accuracy (12–15). Currently, AI algorithms have proven successful in processing clinical image data, obtaining desired prediction results (16–18). For example, a two-stage convolutional neural network (CNN) model was proposed to predict the occurrence of myocardial infarction and localize the site of infarction based on vectorcardiogram signals (19). However, further research is needed to process clinical non-image data using machine learning.

In this study, we aimed to develop an artificial intelligence prediction model to regressively identify the recurrence risk of breast cancer patients after operation. We used SVM, XGBoost, and LSTM algorithms to integrate the histopathological and molecular characteristics of tumors in patients' EHRs. We also validated the model's performance in predicting risk categories for patients who received neoadjuvant and postoperative chemotherapy or postoperative chemotherapy alone, which can provide a precise assessment for personalized medicine for cancer patients. Our study made the following important contributions:

- Developed models based on three AI algorithms (SVM, XGBoost, and LSTM) that accurately predicted the recurrence risk of postoperative breast cancer patients.
- Provided a suitable model for recurrence risk prediction that reflects the chemotherapy resistance of postoperative patients.
- Our LSTM model approximately evaluated the actual benefit of patients receiving neoadjuvant chemotherapy.
- Predicted recurrence risk by the LSTM model, accurately reflecting the prognosis of postoperative breast cancer patients.

Abbreviations: AI, artificial intelligence; AJCC, American Joint Committee on Cancer; ANN, artificial neural networks; AUC, area under the curve; AUC-PR, area under the precision-recall curve; CACA-CBCS, Chinese Anti-Cancer Association, Committee of Breast Cancer Society; EHR, electronic health record; EHRs, electronic health record systems; ER, estrogen receptor; FISH, fluorescence *in situ* hybridization; GLOBOCAN, Global Cancer Observatory; HER2, human epidermal growth factor receptor 2; IDFS, invasive disease-free survival; LSTM, long short-term memory networks; NLP, natural language processing; PR, progesterone receptor; ROC, receiver operating characteristic curve; SD, standard deviation; SVM, support vector machines; TTP, time to progression; UMLS, Unified Medical Language System; XGBoost, Extreme Gradient Boosting.

Methods

Clinicopathological data of breast cancer patients

This retrospective study was designed to predict the risk of breast cancer patients who underwent surgery through automated models. The overall methodology of this study is illustrated in

Figure 1. A total of 1,962 patients with breast cancer were recruited from the Third Affiliated Hospital of Harbin Medical University from 11/05/2011, to 29/12/2019. There were 121 (6.1%) patients initially excluded because of incomplete pathological examination results or lack of clinical notes. Eventually, 1,841 patients were included in this retrospective analysis. A total of 432 patients underwent different treatment regimens following surgery and had complete treatment information, including radiotherapy, chemotherapy alone, combination therapy, endocrine therapy, and targeted therapy. Completed follow-up information of postoperative patients was collected, containing the surveillance of contralateral breast cancer, lymph node metastases, distant organ metastases, and other relevant monitoring. All study procedures were thoroughly reviewed and received ethical approval from the Harbin Medical University Ethics Committee. Informed written consent was obtained from each participant prior to their involvement in the study. A detailed description of the patient characteristics is found in Supplemental Table 1.

Data parsing and feature extraction

Data preprocessing plays an important role in the application of machine learning (20). Since medical professionals have multiple expressions in medical reports, we first broke each note into blocks and standardized the reporting format, mainly regarding its clinical concepts and attributes. More details are explained in the Supplement Data. We further used natural language processing (NLP) based on the regular expression (regex) in Python to extract all key terms from EHRs (21). The regular expression can quickly analyze large volumes of textual information and has a specialized syntax. We compiled the regular expression pattern for each feature according to this specified syntax, thus accurately matching specific strings (22). An example below shows the feature extraction process:

re.compile (r'ER\([\+\-].?\)|ER\([\+\-].*?\)',re.I)*

- re.compile: this regular expression was given to return every line in which the term "ER(+)" or "ER(-)" is present. Parentheses were probably performed with Chinese and English format in our data.
- re.I: re.IGNORECASE, this function was given to return values treated as case-insensitive.
- re.findall() function was next given to return all the matched strings "[-]" or "[+]" in the form of a list of numeric labels "0" or "1."

In addition, NegEx was used to identify whether a term had been negated, effectively rectifying false-positive cases (23). For instance, "lymph nodes are not enlarged," "lymph node-negative," and "no evidence of lymphovascular invasion" were considered negative. After feature extraction, we combined all the features and created a dataset. The output values of all samples were displayed on the label with "=1" to match successfully; else, it was "=0" (Table 1). The missing values in our raw data were filled in "=0." Eventually, the accuracy of feature extraction was estimated using the actual values in the original text snippets (24). Correct extraction was considered true positive (TP) when the extracted values matched the actual values. A classification for the module was regarded as false positive (FP) when the extracted values did not match the actual values. Missed entities

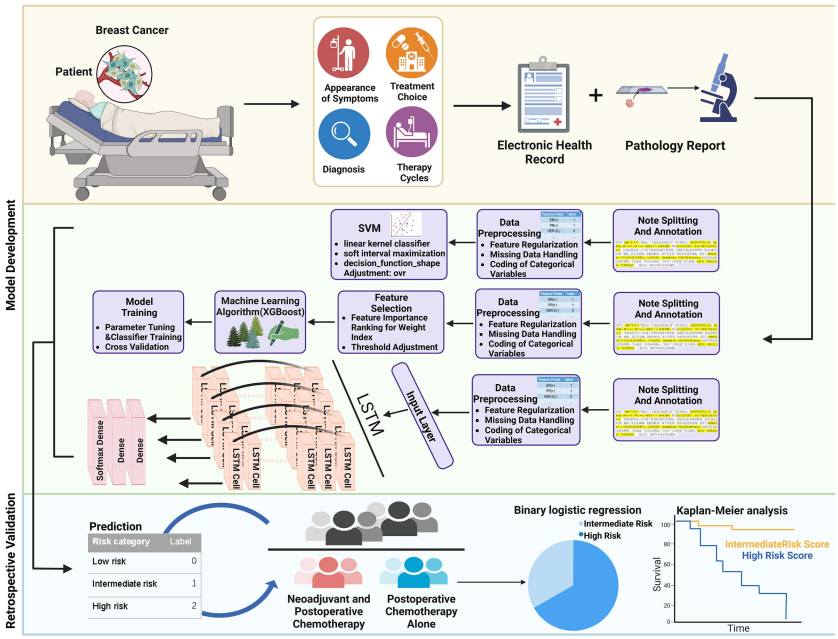

FIGURE 1
Overall workflow of the study. Histopathological features of breast cancer were first extracted and annotated by retrospective retrieving EHR data of breast cancer patients. The preprocessed information was next generated as a feature set, and models were trained to predict the recurrence risk of patients. The model was further validated in patients who received neoadjuvant and postoperative chemotherapy or postoperative chemotherapy alone.

were considered false negative (FN) when actual values were available, but no extracted values were reported. It was regarded as a true negative (TN) when no extracted values were produced and there were no actual values. Supplementary Table 2 shows the confusion matrix for evaluated extraction.

Model prediction and evaluation

The recurrence risk of postoperative breast cancer patients was according to the clinical guidelines for the diagnosis and treatment of Breast cancer in 2021 of Chinese Anti-Cancer Association, Committee of Breast Cancer Society (CACA-CBCS) (Supplemental Table 3) (25). It has performed an important premise that Chinese clinicians base on to comprehensively assess and formulate treatment regimens.

Each prediction model was implemented through the Scikit-learn library in Python. First, the dataset was loaded into the Pandas dataframe and split into a training set (70%) and a test (30%) set with the train_test_split function. In order to avoid extreme values, the fillna() function was executed to fill the vacant values with default values and scale numerical variables for range adjustment.

SVM is a supervised learning algorithm commonly employed in binary classification and regression problems. The basic principle of SVM is to identify a decision boundary so that samples can be separated from different classes (26). In this study, the sklearn.svm.SVC function was adopted to solve the three classification problems. The linear kernel was first selected to linearly classify the training set due to the significantly larger feature size than the sample size (27). Since our data are linearly non-separable, slack variables were employed during the training process to improve the generalization ability of the model by allowing some sample points to be misclassified. Additionally, the decision hyperplane was determined by soft margin maximization and dual problem settlement. The application of multiclass classification utilized a one-*vs*-rest voting strategy, which means that three binary classifiers are trained (28). Finally, samples from the test set were predicted separately and the category with the highest probability was subsequently assigned as the final prediction.

The XGBoost model contains K base learners in which each learner predicts the X_i outcome of the i-th input and then acquires the final classification result by pooling each output $f_K\ (X_i)$ (11). The xgb.XGBClassifier function was adopted to build the model based on a set of relevant parameters such as learning rate, number of trees, and gamma. The grid search strategy was applied from the Sklearn interface to obtain the best-optimized hyperparameters, which optimizes the model's performance and avoids overfitting issues (29). Next, the XGBoost model was trained using the determined parameters and 10-fold cross-validation (30). The most important features that were taken into consideration were as follows: distant organ metastasis, lymph node metastasis (including the number of lymph node metastases), HER-2, ER, PR, and Ki-67 expression; pathology grade; menopausal status; age; and lympho-vascular invasion. Eventually, values were predicted for the test set and evaluated by the module to obtain the reliability of the XGBoost model (31).

LSTM simulates the memory storage capacity of our brain, which develops novel artificial intelligence algorithms. Compared with traditional neural network algorithms, LSTM can precisely deal with more complex problems related to time series or sequential data (32, 33). In this study, the LSTM model was constructed in Keras. After learning meaningful features, dense layers were used to map features from the high-dimensional data space to a low-dimension representation space and finally become a column vector, in which the number of columns is the same as risk categories (34). Specifically, the first column corresponded to the low risk with "class 0," the second column corresponded to the intermediate risk with "class 1," and the third column corresponded to the high risk with "class 2." Each patient would be obtained a column vector with a sum of 1 through the softmax_layer. For example, the predicted result for one patient was shown [0.2,0.7,0.1]. A predicted value of 0.2 represented the probability of class "0," 0.7 represented the probability of class "1," and 0.1 represented the probability of class "2." This column vector indicated that this patient was finally classified as the maximum value of the predicted label "class-1" (intermediate-risk). Moreover, backpropagation was utilized to optimize the parameters of this model, thus minimizing the loss function (35). Feature units were randomly dropped through dropout layers during each feedforward training to avoid overfitting issues and obtain a generalization model.

To determine the favorable model, the performance of each model was compared through the receiver operating characteristic curve (ROC) and the area under the curve (AUC). Since our dataset has an imbalanced distribution of samples, consisting of disparate sample sizes in each class. Precision (positive predictive value)–recall (sensitivity) curves were also applied as indicators to further assess each model's performance (36). Other important metrics for evaluation include accuracy, F1 score, macro-average, micro-average, and weight-average. A further explanation of these indicators is provided in the Supplement Data.

Statistical analysis in patients

We divided the 85 patients treated with chemotherapy alone after surgery into chemo-sensitive and chemo-resistant groups based on each patient's response to chemotherapy. The inclusion criteria of chemotherapeutic resistance are as follows (37, 38): (1) An increase in tumor volume after postoperative chemotherapy was observed using B-ultrasound and MRI; (2) sustained increases in tumor marker levels and clinical symptoms did not relieve; (3) and patients were confirmed as having progressive disease (PD) according to Response Evaluation Criteria in Solid Tumors (RECIST version 1.1). Chemotherapy resistance is considered when one or all of the criteria are met. In order to retrospectively validate the predictive effectiveness of our model, we next used a binary logistic regression approach with chemotherapy resistance as the dependent variable and the risk categories predicted by our model as the covariate (39).

For patients treated with neoadjuvant chemotherapy, the endpoint was time to progression (TTP) because a death event was not observed at the cutoff in this study. TTP was defined as the date from registration to invalid treatment or disease progression (40, 41). For subgroups only undergoing postoperative chemotherapy, the endpoint of interest was set as invasive

TABLE 1 Feature extracted labels and descriptions.

Feature names	Feature descriptions	Illustrative example
Patients	Patient ID	616402
Age	Years	51
Menopausal status	Pre = 0 Post = 1	1
ER	Estrogen receptor-positive = 1 Estrogen receptor-negative = 0	1
PR	Progesterone receptor-positive = 1 Progesterone receptor-negative = 0	1
HER2	HER2/neu gene overexpressed or amplified = 1 HER2/neu gene neither overexpressed nor amplified = 0	0
Tumor size	Pathological tumor size ≤2cm = 0 Pathological tumor size >2 cm = 1	0
LNM	Positive lymph node metastasis = 1 Negative lymph node metastasis = 0	0
Number of LNM	The number of lymph node metastases	
G1	Pathology grade I = 1 Pathology grade II, pathology grade III = 0	1
G2	Pathology grade II = 1 Pathology grade I, pathology grade III = 0	0
G3	Pathology grade III = 1 Pathology grade I, pathology grade II = 0	0
LVI	Lympho-vascular invasion (+) = 1 Lympho-vascular invasion (-) = 0	0
Ki-67 (%)	The median pathology of Ki-67 proliferative index	5
Distant organ metastasis	Distant organ metastasis = 1 Non-distant organ metastasis = 0	0
Label	Low risk = 0 Intermediate risk = 1 High risk = 2	0

disease-free survival (iDFS). iDFS is calculated as the time interval from the date of registration to the first recurrence of breast cancer, the development of contralateral primary breast cancer, or death from any cause (42).

Kaplan–Meier analysis and the log-rank test were used to assess survival outcomes in groups treated with neoadjuvant and postoperative chemotherapy as well as postoperative chemotherapy alone. All statistical analyses were implemented with the R software 3.5.0 (https://www.r-project.org/); a P value <0.05 was considered statistically significant.

Results

Training and test cohorts conducted

The included cohorts were randomly divided into training and test cohorts according to the ratio of 7(n = 1,289):3(n = 552) (Table 2) (43, 44). The validation set was considered a part of the training cohort to fine-tune the hyperparameters in our models. Each group of information was evenly distributed without bias. Table 2 presents the characteristics of patients. Valuable information in EHRs was first segmented and annotated, including integrated pathological and clinical information from encounter notes and progress notes. Text snippets were further processed using feature extraction methods to extract specific string fields (45). The extractor achieved 95% accuracy, and each string was then matched against the numeric label "0" or "1"; all matched features of each patient were aggregated together to form a large dataset, which simplifies the learning process. This transformation process involved converting complex multiple input variables into a more manageable format, which greatly improved the classification performance of our model (46). The standards for automatic extraction are shown in the methods.

SVM, XGBoost, and LSTM models predicted the recurrence risk of postoperative breast cancer patients

After model development with the training subset, test samples were uploaded to predict recurrence risk, and this multi-classification task was conducted *via* a one-*vs*-the-rest method.

TABLE 2 Cohort characteristics for 30% train/70% test experiments in breast cancer patients.

Characteristic	Training set	Test set
Number of patients	1,289	552
Gender, %Female	1,282 (99.5)	549 (99.5)
Gender, %Male	7 (0.5)	3 (0.5)
Age, no. (%)		
<35	12 (0.9)	8 (1.4)
≥35	1,277 (99.1)	544 (98.6)
Menopausal status		
Pre	298	109
Post	984	440
Molecular subtypes		
Luminal A/luminal B	609	248
HER2+	429	178
Triple negative	251	126
Histology		
Invasive ductal carcinoma	1,067	443
Invasive lobular carcinoma	53	23
Mixed (IDC and ILC)	48	16
DCIS/LCIS	86	58
Other types	35	12
Recurrence risk assessment		
Low-risk	102	86
Intermediate-risk	758	283
High-risk	429	183

IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; DCIS, ductal carcinoma in situ; LCIS, lobular carcinoma in situ; F, female; HER2, human epithelial growth factor receptor-2.

Specifically, when one category was correctly predicted by the model, the remaining categories were considered negative (47), thus generating a confusion matrix for each category (Figure 2). We computed the evaluation metrics of each category based on the confusion matrix, such as accuracy, precision, recall, and the area under the receiver-operating characteristic curve (ROC-AUC) (Table 3 and Figure 3). In order to further compare the effectiveness of models, we averaged (macro-average, F1 score) and weighted (micro-average, weighted-average) the evaluation indicators of each category (Table 3) (47, 48). Subsequently, we draw the ROC curve for each prediction category with the true positive rate (TPR) as the abscissa and the false positive rate (FPR) as the ordinate and explained the achievement of each model using a micro-average ROC curve (Figures 3A, C, E). The AUC values of the micro-average ROC curve corresponding to SVM, XGBoost, and LSTM were 0.92 ± 0.06, 0.97 ± 0.03, and 0.98 ± 0.01 (Figures 3A, C, E). Additionally, the area under the precision–recall curve (AUC-PR) is more suitable for assessing performance metrics on processing imbalanced data compared with the area under the receiver operating characteristic curve (AUC-ROC) (49–51). The SVM generated the smallest micro-average AUC-PR (0.86 ± 0.11), and the LSTM model demonstrated the largest micro-average AUC-PR (0.96 ± 0.02), which indicates that a great number of patients were correctly labeled (Figures 3B, D, F). Overall, the LSTM model accomplished superior performance on the test set, with a micro-averaged AUC-PR that represents an improvement of 10% and 3% compared with SVM and XGBoost. The LSTM model manifested a significantly higher accuracy (0.89), F1 score, macro-F1 score (0.87), and weighted-F1 (0.89) (Table 3).

Breast cancer patients at high recurrence risk are more likely to be resistant to chemotherapy after surgery

Chemotherapy resistance is the most crucial reason for recurrence of breast cancer patients after surgery (52). In order to exclude the influence of other treatment options on the effect of chemotherapy, patients who received chemotherapy alone were included in the experiment. A binary logistic regression analysis was executed to identify the association between model-based predicted recurrence risk and chemotherapy resistance in breast cancer. The inclusion criteria for chemotherapy resistance in this study are described in the methods. A total of 432 patients received postoperative treatment, and 85 (20%) patients underwent chemotherapy alone, which included DNA-damaging drugs such as anthracyclines and platinum and microtubule-targeting drugs like paclitaxel. There were 37 patients classified as high-risk by the LSTM model, 32 of which (86%) were chemotherapy resistant. Among the 46 intermediate-risk patients predicted by the LSTM model, 29 (63%) patients were chemotherapy resistant (Table 4). The results of binary logistic regression showed that the probability of DNA-damaging drug resistance in high-risk patients predicted by the LSTM model was 4.062 times more than in intermediate-risk patients ($P < 0.05$; Figure 4A). Meanwhile, the high-risk patients predicted by the LSTM model were more likely to be resistant to microtubule-targeted drugs than the intermediate-risk patients (high-risk: intermediate-risk = 5.667: 1; $P < 0.05$; Figure 4A). These results suggest that high-risk patients predicted by our model are more resistant to chemotherapy drugs after surgery and likely to perform more insensitively to paclitaxel. Consistent results were observed in the SVM and XGBoost models, but the P values are not significant (Figures 4B, C). We did not include the low-risk patients because the number of low-risk samples was insufficient to meet the minimum sample size (n = 10) required for binary logistic regression analysis.

Our model can predict the neoadjuvant chemotherapy benefits and the survival of patients

Neoadjuvant therapy plays an important role in the clinical practice of systemic treatment for breast cancer patients (53).

10.3389/fonc.2023.1117420

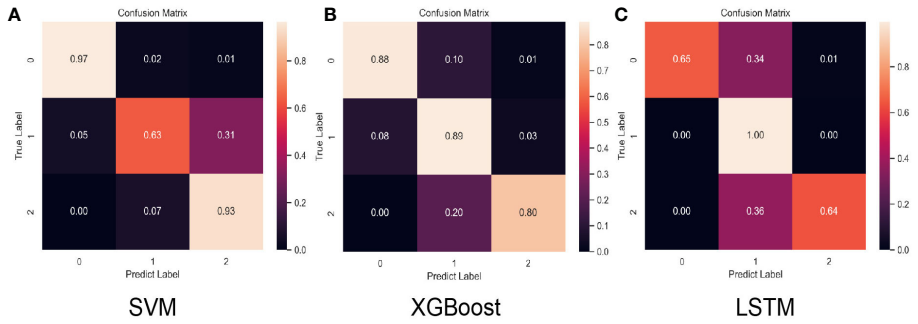

FIGURE 2
Normalized confusion matrix for the test set of each model. "Class 0," "class 1," and "class 2" correspond to low-risk, intermediate-risk, and high-risk categories. **(A)** SVM confusion matrix, **(B)** XGBoost confusion matrix, **(C)** LSTM confusion matrix.

Nevertheless, recent research has reported that neoadjuvant chemotherapy is not necessarily beneficial for patient survival. Patients who were refractory to neoadjuvant treatment can result in a higher local recurrence rate after surgery (54, 55). Among our subgroups treated with neoadjuvant chemotherapy, 52 and 72 patients were predicted to be intermediate and high risk by LSTM, respectively (Table 5). Contrary to our anticipated outcome, the results indicate that the majority of patients who received neoadjuvant chemotherapy did not exhibit a low-risk profile as we had expected. Moreover, 43 and 23 patients treated with the postoperative chemotherapy alone were predicted as intermediate risk and high risk by LSTM. These results indicated that not all breast cancer patients should receive neoadjuvant chemotherapy before surgery. Our predictive model can be utilized to evaluate the benefit of patients receiving neoadjuvant chemotherapy.

Data were next manually extracted on time to disease progression (TTP), which was considered a reliable surrogate endpoint in advanced cancer with medical therapy (Lee, Jang, Lee, Cho, Lee, Yu, Kim, Yoon, Kim, Han, Oh, Im and Kim 2016). For patients administered neoadjuvant chemotherapy, the intermediate-risk operated patients predicted by the LSTM model was shown to have a longer TTP than the high-risk ones ($P < 0.05$;

TABLE 3 Comparison of test set prediction performance between the models.

	Precision	Recall	F1 score	Accuracy
SVM				
Low-risk	0.85	0.97	0.90	0.78
Intermediate-risk	0.93	0.63	0.75	
High-risk	0.64	0.93	0.76	
Macro avg	0.81	0.84	0.81	
Weighted avg	0.82	0.78	0.78	
XGBoost				
Low-risk	0.76	0.88	0.82	0.86
Intermediate-risk	0.85	0.89	0.87	
High-risk	0.94	0.80	0.86	
Macro avg	0.85	0.86	0.85	
Weighted avg	0.86	0.86	0.86	
LSTM				
Low-risk	1	0.65	0.79	0.89
Intermediate-risk	0.83	1	0.91	
High-risk	0.99	0.84	0.91	
Macro avg	0.94	0.83	0.87	
Weighted avg	0.91	0.89	0.89	

10.3389/fonc.2023.1117420

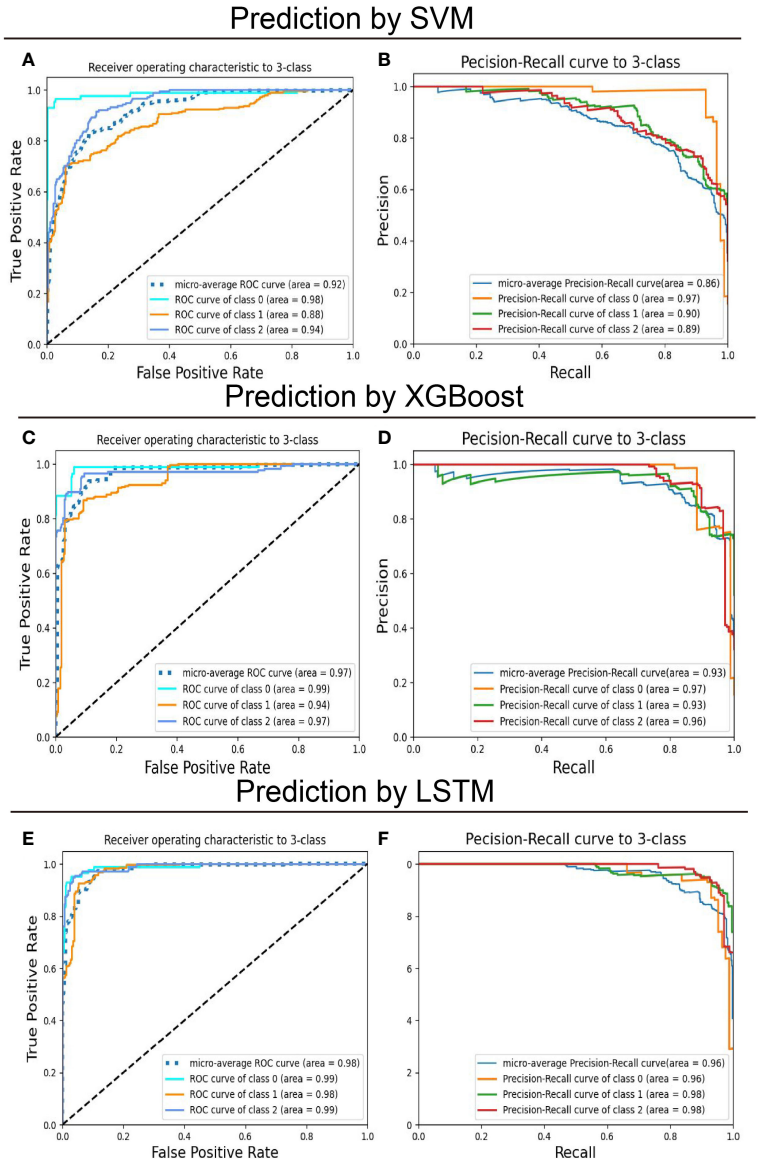

FIGURE 3
Predictive performance of models on the training set for multiclassification of breast cancer patients. The support vector machines (SVM), extreme gradient boosting (XGBoost), and long short-term memory (LSTM) recurrent neural network models were trained to classify patients with operated breast cancer from the feature label values. **(A, C, E)** Receiver-operating characteristics (ROC) curve and **(B, D, F)** Precision-recall (PR) curve for the test set was shown to quantify the performance of models. "Class 0," "class 1," and "class 2" correspond to low-risk, intermediate-risk, and high-risk categories.

Figure 5A). We compared invasive disease-free survival (IDFS) in the groups that received only postoperative chemotherapy and found that the high-risk patients acquired poorer IDFS than the intermediate-risk ones ($P < 0.05$; Figure 5B). Compared with intermediate-risk or high-risk, the low-risk sample size was insufficient to create reliable estimates. However, low-risk patients actually had better outcomes according to their clinical information. Therefore, our model can accurately predict the prognosis of breast cancer patients before treatment and suggest that clinicians provide the most appropriate treatment regimen for patients, such as whether to administrate patients with neoadjuvant chemotherapy or postoperative chemotherapy.

Discussion

In this study, the advantages and limitations of our proposed model are as follows: (i) All models can seamlessly classify from labeled data with an accuracy of over 75%. (ii) The linear SVM model generates a good non-linear mapping between input and output variables. It has good robustness and appears to have no effect on the model when non-supported vector samples are added and removed, thus avoiding the problems of leaf node selection in XGBoost and dimension disaster in LSTM. (iii) The XGBoost model excited more parameters and performed more accurately than SVM. It illustrated a white box compared with ANN so that

TABLE 4 Predictive performance of the LSTM model for postoperative breast cancer patients treated with chemotherapy alone.

Number of patients	Recurrence risk assessment Low-risk (AUC ± SD)	Recurrence risk assessment Intermediate-risk (AUC ± SD)	Recurrence risk assessment High-risk (AUC ± SD)
Chemo-sensitive	2 (0.92 ± 0.03)	17 (0.87 ± 0.04)	5 (0.85 ± 0.08)
Chemo-resistant	0	29 (0.84 ± 0.07)	32 (0.86 ± 0.11)

The model's performance was assessed through the area under the curve (AUC) ± standard deviation (SD).

the model's effectiveness can be intuitively evaluated. Moreover, the XGBoost model has presorted features based on the parameters before training, which were repeatedly utilized in subsequent iterations, significantly reducing the computation. (iv) LSTM realized the highest accuracy among all models, attributed to the continuous optimization of gradient descent and backpropagation. (v) The high recurrence risk predicted by the LSTM model was consistent with the chemotherapy resistance and the worse prognosis of postoperative patients, which corresponded to the actual situation. (vi) The SVM algorithm is less sensitive to the handling of missing data. Clearly, vacant values were filled with the default value "0" during data preprocessing, which affects the linear separability in the feature space of SVM. Nevertheless, the XGBoost algorithm tries different methods at each node and identifies the best method to handle when missing data are encountered. LSTMs can learn complex correlations between features, including further details in default values. (vii) The model uses only a single type of input information that converts textual clinical reports into labeled values. Once new variables emerge, we will manually develop and validate a new set of regular expressions for each specific task.

We established machine learning algorithms capable of extracting patient classification information from unstructured clinical notes. Benefitting from the application of technologies and frameworks of machine learning, our models for screening diagnostics with low-cost burden were favorable (56). However, several unavoidable challenges with machine learning were posed. First, the data annotation and processing were complicated. In order to achieve data collection and annotation with high precision, including the term standardization of biological features, the variability of descriptive words, and the presence of negative phrases, we searched for each key term and encoded it with category encoders through feature engineering and natural language processing. High accuracy was achieved eventually for each feature of information abstracting. Secondly, for high-dimensional scene data exploration (such as medical time-series data), the XGBoost algorithm cannot effectively eliminate noise variables (57). Therefore, we conducted a grid search to determine the algorithms of optimal dimensionality reduction and added randomness to improve robustness (58). Additionally, an increasing fraction of the training time in the LSTM model would

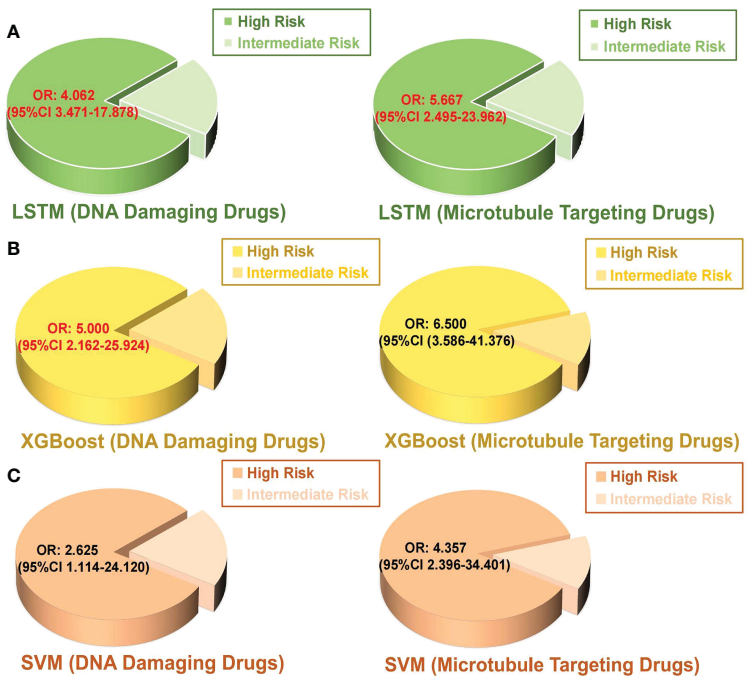

FIGURE 4
Binary logistic regression was performed to analyze the relationship between the predicted intermediate risk and high risk by each model and chemotherapy resistance in postoperative breast cancer patients. **(A)** LSTM, **(B)** XGBoost, **(C)** SVM. OR value: odd ratio; The red color indicates a statistically significant correlation $P < 0.05$.

TABLE 5 Predictive performance of the LSTM model for breast cancer patients treated with neoadjuvant and postoperative chemotherapy or postoperative chemotherapy alone.

Number of patients	Recurrence risk assessment Low-risk (AUC ± SD)	Recurrence risk assessment Intermediate-risk (AUC ± SD)	Recurrence risk assessment High-risk (AUC ± SD)
Neoadjuvant and postoperative chemotherapy	3 (0.93 ± 0.01)	52 (0.91 ± 0.03)	72 (0.95 ± 0.03)
Postoperative chemotherapy alone	2 (0.89 ± 0.04)	43 (0.87 ± 0.02)	23 (0.89 ± 0.01)

For patients treated with neoadjuvant and postoperative chemotherapy or postoperative chemotherapy alone, the model was trained to extract postoperative information to classify "high-risk," "intermediate-risk," and "low-risk" labels. The model's performance was assessed through the area under the curve (AUC) ± standard deviation (SD).

reduce the number of iterations within the same total training time (59). We utilized forward calculation and backpropagation to continuously adjust the parameters for extracting the optimal features. Therefore, we provided a reproducible predicted tool to predict the recurrence risk of breast cancer patients after surgery.

To further guide clinical practice, our models maintained their performance in reflecting patient tolerance to chemotherapy drugs. We verified that high-risk patients tend to be more resistant to DNA damage and microtubule inhibition drugs than intermediate-risk patients. This result provides a basis for the clinical treatment application of different drugs to postoperative breast cancer patients. Chemotherapy resistance is not only an important risk factor for cancer recurrence but also a major cause of poor patient outcomes (52). Meanwhile, our models also validated the prognosis of patients who underwent neoadjuvant chemotherapy and postoperative chemotherapy. Since the linkages between EHR data and death registries were rare, we used TTP or IDFS as surrogate endpoints to assess differences in survival outcomes of predicted categories. Our approach highlighted the importance of estimating the recurrence risk after neoadjuvant chemotherapy, indicating whether patients routinely receive preoperative chemotherapy is worth thought-provoking (60, 61). Although patients classified as low-risk were predicted in our model, the recurrence was not statistically significant compared with the other two groups because of the rare number of samples.

Previous studies have applied natural language processing to abstract biological factors from medical records to predict breast cancer staging based on the American Joint Committee on Cancer (AJCC) staging manual (62). In 2020, researchers also implemented artificial neural networks to predict breast cancer prognosis by selecting crucial survival factors, including tumor size, tumor staging, lymph node metastasis, and other related variables (63). Moreover, deep learning has shown promise in predicting breast cancer risk rates by extracting factors such as age, race, and menstrual history (64). In contrast, our approach significantly solved the bottleneck of extracting outcomes from a great number of clinical texts and achieved effective feature extraction in different scenes. Additionally, those included studies were predominantly conducted in the United States or Europe, but the data for Asian breast cancer patients remained unknown. Breast cancer incidence is strongly correlated with variations in geographic distribution (65, 66). Because of differences in people's diets and lifestyles, breast cancer is highly prevalent in the alpine region (67), such as the northeast of China. An accurate assessment of patients' recurrence risk before tailored individual treatment plans can provide valuable guidance on improving patient outcomes. Our studies contribute to the development of screening strategies for breast cancer in the Asian population.

In conclusion, we developed AI-based models that integrate histopathological features of breast cancer and clinical information

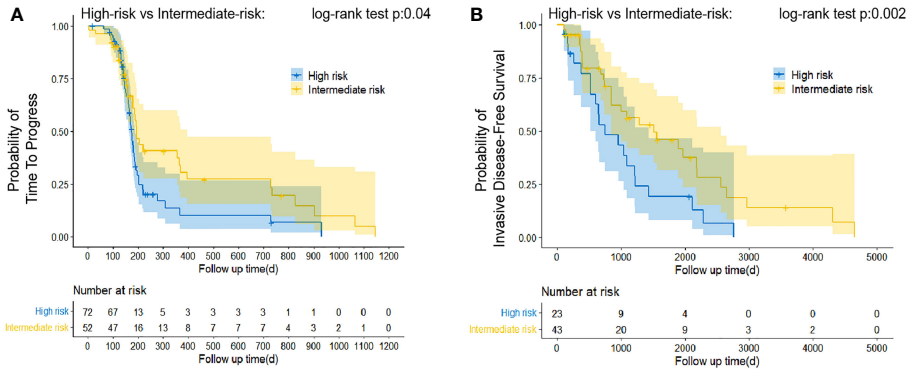

FIGURE 5
Estimation of relative survival in classified patients treated with neoadjuvant therapy and postoperative chemotherapy alone by Kaplan–Meier curve analysis. Patients predicted to be classified as "Intermediate-risk" presented favorable TTP **(A)** and iDFS **(B)** than that patient identified as "high-risk." The log-rank test was appropriate to assess performance.

from preprocessed clinical notes to predict the recurrence risk of postoperative breast cancer patients. The performance and generalizability of our model have emphasized the potential application in the estimation of recurrence risk in breast cancer patients.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of Harbin Medical University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

WY offered main direction and significant guidance of this manuscript. LZ, LL, DC and HL drafted the manuscript and illustrated the figures for the manuscript. LL provided the clinical data. YX and HB helped with the data analyzed. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the Heilongjiang Province Postdoctoral Scientific Research Developmental Fund (Grant Number: LBH-Q20046 to WY) and the Natural Science Foundation of Heilongjiang Province of China to WY (Grant Number: LH2022H008).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2023.1117420/full#supplementary-material

References

1. Huang J, Chan PS, Lok V, Chen X, Ding H, Jin Y, et al. Global incidence and mortality of breast cancer: A trend analysis. *Aging (Albany NY)* (2021) 13:5748–803. doi: 10.18632/aging.202502

2. Cao W, Chen HD, Yu YW, Li N, Chen WQ. Changing profiles of cancer burden worldwide and in China: A secondary analysis of the global cancer statistics 2020. *Chin Med J (Engl)* (2021) 134:783–91. doi: 10.1097/CM9.0000000000001474

3. Chlebowski RT. Improving breast cancer risk assessment versus implementing breast cancer prevention. *J Clin Oncol* (2017) 35:702–4. doi: 10.1200/JCO.2016.70.9386

4. Shimizu H, Nakayama KI. Artificial intelligence in oncology. *Cancer Sci* (2020) 111:1452–60. doi: 10.1111/cas.14377

5. Sultan AS, Elgharib MA, Tavares T, Jessri M, Basile JR. The use of artificial intelligence, machine learning and deep learning in oncologic histopathology. *J Oral Pathol Med* (2020) 49:849–56. doi: 10.1111/jop.13042

6. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* (2021) 13:152. doi: 10.1186/s13073-021-00968-x

7. Jiang X, Xu C. Deep learning and machine learning with grid search to predict later occurrence of breast cancer metastasis using clinical data. *J Clin Med* (2022) 29:11. doi: 10.3390/jcm11195772

8. Evans RS. Electronic health records: Then, now, and in the future. *Yearb Med Inform* (2016) Suppl 1:S48–61. doi: 10.15265/IYS-2016-s006

9. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *J Am Med Inform Assoc* (2019) 26:364–79. doi: 10.1093/jamia/ocy173

10. Tripathy RK, Mahanta S, Paul S. Artificial intelligence-based classification of breast cancer using cellular images. *RSC Adv* (2014) 4:9349–55. doi:10.1039/C3RA47489E

11. Chen T, Guestrin C. (2016). *XGBoost: A Scalable Tree Boosting System*. New York,USA: Association for Computing Machinery.

12. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FEJr., et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* (1997) 79:857–62. doi: 10.1002/(SICI)1097-0142(19970215)79:4<857::AID-CNCR24>3.0.CO;2-Y

13. Schmidhuber J. Deep learning in neural networks: An overview. *Neural Netw* (2015) 61:85–117. doi: 10.1016/j.neunet.2014.09.003

14. Tsai YC, Lai SH, Ho CJ, Wu FM, Henrickson L, Wei CC, et al. High accuracy respiration and heart rate detection based on artificial neural network regression. *Annu Int Conf IEEE Eng Med Biol Soc* (2020) 2020:232–5. doi: 10.1109/EMBC44109.2020.9175161

15. Zdolsek G, Chen Y, Bogl HP, Wang C, Woisetschlager M, Schilcher J. Deep neural networks with promising diagnostic accuracy for the classification of atypical femoral fractures. *Acta Orthop* (2021) 92:394–400. doi: 10.1080/17453674.2021.1891512

16. Dai G, Zhang X, Liu W, Li Z, Wang G, Liu Y, et al. Analysis of EPID transmission fluence maps using machine learning models and CNN for identifying position errors in the treatment of GO patients. *Front Oncol* (2021) 11:721591. doi: 10.3389/fonc.2021.721591

17. Jiang YQ, Cao SE, Cao S, Chen JN, Wang GY, Shi WQ, et al. Preoperative identification of microvascular invasion in hepatocellular carcinoma by XGBoost and deep learning. *J Cancer Res Clin Oncol* (2021) 147:821–33. doi: 10.1007/s00432-020-03366-9

18. Kumar A, Vatsa A. Untangling classification methods for melanoma skin cancer. *Front Big Data* (2022) 5:848614. doi: 10.3389/fdata.2022.848614

19. Karhade J, Ghosh SK, Gajbhiye P, Tripathy RK, Acharya UR. Multichannel multiscale two-stage convolutional neural network for the detection and localization of myocardial infarction using vectorcardiogram signal. *Applied Sci* (2021) 11:7920. doi: 10.3390/app11177920

20. Malley B, Ramazzotti D, Wu JT. *Data pre-processing. in: Secondary analysis of electronic health records*. Cham (CH): Springer, Cham (2016) p. 115–41.

21. Fu JT, Sholle E, Krichevsky S, Scandura J, Campion TR. Extracting and classifying diagnosis dates from clinical notes: A case study. *J BioMed Inform* (2020) 110:103569. doi: 10.1016/j.jbi.2020.103569

22. Veena G, Hemanth R, Hareesh J. (2019). *Relation extraction in clinical text using NLP based regular expressions*. Kannur, India: Institute of Electrical and Electronics Engineers (IEEE).

23. Mehrabi S, Schmidt CM, Waters JA, Beesley C, Krishnan A, Kesterson J, et al. An efficient pancreatic cyst identification methodology using natural language processing. *Stud Health Technol Inform* (2013) 192:822–6. doi: 10.3233/978-1-61499-289-9-822

24. Kitchenham B, Pfleeger SL, McColl B, Eagan SJ. An empirical study of maintenance and development estimation accuracy. *J Syst Softw* (2002) 64:57–77. doi: 10.1016/S0164-1212(02)00021-3

25. Breast Cancer Expert Committee of China Anti-Cancer A. [Guidelines for clinical diagnosis and treatment of breast cancer in China, (2021 edition)]. zhonghua. *Zhong Liu Za Zhi* (2021) 31:954–1040. doi: 10.3760/cma.j.cn112152-20200817-00747

26. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* (2015) 13:8–17. doi: 10.1016/j.csbj.2014.11.005

27. Jakkula V. Tutorial on support vector machine (svm). *Comput Sci* (2006) 37:3. Available at: https://www.semanticscholar.org/paper/Tutorial-on-Support-Vector-Machine-(-SVM-)-Jakkula/7cc83e98367721bfb908a8f703ef5379042c4bd9.

28. Varpa K, Joutsijoki H, Iltanen K, Juhola M. Applying one-vs-one and one-vs-all classifiers in k-nearest neighbour method and support vector machines to an otoneurological multi-class problem. In: *User centred networked health care*. Tampere, Finland: IOS Press (2011). p. 579–83.

29. Nguyen G, Dlugolinsky S, Bobák M, Tran V, López García Á, Heredia I, et al. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artif Intell Rev* (2019) 52:77–124. doi: 10.1007/s10462-018-09679-z

30. Ramraj S, Uzir N, Sunil R, Banerjee S. Experimenting XGBoost algorithm for prediction and classification of different datasets. *Int J Control Theory and Applications* (2016) 9:651–62. Available at: https://www.researchgate.net/publication/318132203_Experimenting_XGBoost_Algorithm_for_Prediction_and_Classification_of_Different_Datasets.

31. Brownlee J. *XGBoost with python: Gradient boosted trees with XGBoost and scikit-learn*. Greater Melbourne Area: Machine Learning Mastery (2016).

32. Yao K, Cohn T, Vylomova K, Duh K, Dyer C. Depth-gated LSTM. (2015). doi: 10.48550/arXiv.1508.03790

33. Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput* (2019) 31:1235–70. doi: 10.1162/neco_a_01199

34. Liu X, Zeng Z, Wunsch DCII. Memristor-based LSTM network with *in situ* training and its applications. *Neural Netw* (2020) 131:300–11. doi: 10.1016/j.neunet.2020.07.035

35. Hou L, Zhu J, Kwok J, Gao F, Qin T, Liu T-y. Normalization helps training of quantized lstm. (2019) 32:. doi: 10.5555/3454287.3454947

36. Liu Z, Bondell HD. Binormal precision–recall curves for optimal classification of imbalanced data. *Stat Biosci* (2019) 11:141–61. doi: 10.1007/s12561-019-09231-9

37. Jiang YZ, Liu Y, Xiao Y, Hu X, Jiang L, Zuo WJ, et al. Molecular subtyping and genomic profiling expand precision medicine in refractory metastatic triple-negative breast cancer: the FUTURE trial. *Cell Res* (2021) 31:178–86. doi: 10.1038/s41422-020-0375-9

38. Chen Y, Feng X, Yuan Y, Jiang J, Zhang P, Zhang B. Identification of a novel mechanism for reversal of doxorubicin-induced chemotherapy resistance by TXNIP in triple-negative breast cancer *via* promoting reactive oxygen-mediated DNA damage. *Cell Death Dis* (2022) 13:338. doi: 10.1038/s41419-022-04783-z

39. Huang FL. Alternatives to logistic regression models when analyzing cluster randomized trials with binary outcomes. *Prev Sci* (2021) 22:1–10. doi: 10.1007/s11121-021-01228-5

40. Lee DW, Jang MJ, Lee KH, Cho EJ, Lee JH, Yu SJ, et al. TTP as a surrogate endpoint in advanced hepatocellular carcinoma treated with molecular targeted therapy: meta-analysis of randomised controlled trials. *Br J Cancer* (2016) 115:1201–5. doi: 10.1038/bjc.2016.322

41. Saito T, Murotani K. Treatment of death events in the analysis of time to progression. *Ther Innov Regul Sci* (2022) 56:1–3. doi: 10.1007/s43441-021-00343-3

42. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N Engl J Med* (2018) 379:111–21. doi: 10.1056/NEJMoa1804710

43. Gholamy A, Kreinovich V, Kosheleva O. *Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation*. Texas, USA: Computer Sciences (2018).

44. Wagner J, Rapsomaniki MA, Chevrier S, Anzeneder T, Langwieder C, Dykgers A, et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* (2019) 177:1330–1345 e1318. doi: 10.1016/j.cell.2019.03.005

45. Khalid S, Khalil T, Nasreen S. (2014). *A survey of feature selection and feature extraction techniques in machine learning*. London, UK: Institute of Electrical and Electronics Engineers (IEEE).

46. Guyon I, Elisseeff A. An introduction to feature extraction. In: *Feature extraction*. California, USA: Springer (2006). p. 1–25.

47. Murphy KP. *Machine learning: A probabilistic perspective*. Massachusetts, USA: MIT press (2012).

48. Pillai I, Fumera G, Roli F. (2012). *F-measure optimisation in multi-label classifiers*. Tsukuba, Japan: Institute of Electrical and Electronics Engineers (IEEE).

49. Subbe CP, Slater A, Menon D, Gemmell L. Validation of physiological scoring systems in the accident and emergency department. *Emerg Med J* (2006) 23:841–5. doi: 10.1136/emj.2006.035816

50. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* (2016) 44:368–74. doi: 10.1097/CCM.0000000000001571

51. Ferrari D, Milic J, Tonelli R, Ghinelli F, Meschiari M, Volpi S, et al. Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia-challenges, strengths, and opportunities in a global health emergency. *PLos One* (2020) 15:e0239172. doi: 10.1371/journal.pone.0239172

52. Prihantono, Faruk M. Breast cancer resistance to chemotherapy: When should we suspect it and how can we prevent it? *Ann Med Surg (Lond)* (2021) 70:102793. doi: 10.1016/j.amsu.2021.102793

53. Montemurro F, Nuzzolese I, Ponzone R. Neoadjuvant or adjuvant chemotherapy in early breast cancer? *Expert Opin Pharmacother* (2020) 21:1071–82. doi: 10.1080/14656566.2020.1746273

54. Vaidya JS, Massarut S, Vaidya HJ, Alexander EC, Richards T, Caris JA, et al. Rethinking neoadjuvant chemotherapy for breast cancer. *BMJ* (2018) 360:j5913. doi: 10.1136/bmj.j5913

55. Asaoka M, Gandhi S, Ishikawa T, Takabe K. Neoadjuvant chemotherapy for breast cancer: Past, present, and future. *Breast Cancer (Auckl)* (2020) 14:1178223420980377. doi: 10.1177/1178223420980377

56. Aliabadi A, Sheikhtaheri A, Ansari H. Electronic health record-based disease surveillance systems: A systematic literature review on challenges and solutions. *J Am Med Inform Assoc* (2020) 27:1977–86. doi: 10.1093/jamia/ocaa186

57. Euh S, Lee H, Kim D, Hwang D. Comparative analysis of low-dimensional features and tree-based ensembles for malware detection systems. *Institute of Electrical and Electronics Engineers (IEEE)* (2020) 8:76796–808. doi: 10.1109/ACCESS.2020.2986014

58. Paleczek A, Grochala D, Rydosz A. Artificial breath classification using XGBoost algorithm for diabetes detection. *Sensors (Basel)* (2021) 18:21. doi: 10.3390/s21124187

59. You Y, Hseu J, Ying C, Demmel J, Keutzer K, Hsieh C-J. (2019). *Large-Batch training for LSTM and beyond*. California, USA. doi: 10.48550/arXiv.1901.08256

60. Duchesneau ED, An SJ, Strassle PD, Reeder-Hayes KE, Gallagher KK, Ollila DW, et al. Sociodemographic and clinical predictors of neoadjuvant chemotherapy in cT1-T2/N0 HER2-amplified breast cancer. *Ann Surg Oncol* (2022) 29:3051–61. doi: 10.1245/s10434-021-11260-y

61. Fujita N, Fujita K, Kim SJ, Iguchi C, Nomura T, Aono T, et al. Response-guided omission of anthracycline in patients with HER2-positive early breast cancer treated with neoadjuvant taxane and trastuzumab: 5-year follow-up of prognostic study using propensity score matching. *Oncology* (2022) 100:257–66. doi: 10.1159/000522384

62. Deshmukh PR, Phalnikar R. Information extraction for prognostic stage prediction from breast cancer medical records using NLP and ML. *Med Biol Eng Comput* (2021) 59:1751–72. doi: 10.1007/s11517-021-02399-7

63. Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, et al. Machine learning techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer Med* (2020) 9:3234–43. doi: 10.1002/cam4.2811

64. Stark GF, Hart GR, Nartowt BJ, Deng J. Predicting breast cancer risk using personal health data and machine learning models. *PLos One* (2019) 14:e0226765. doi: 10.1371/journal.pone.0226765

65. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2021) 71:209–49. doi: 10.3322/caac.21660

66. Su H, Li X, Lv Y, Qiu X. Breast cancer epidemiology and survival analysis of shenyang in northeast China: A population-based study from 2008 to 2017. *Breast J* (2022) 282:6168832. doi: 10.1155/2022/6168832

67. Liu B, Lao X, Feng Y, Liu J, Jiao M, Zhao M, et al. Cancer prevalence among the rural poverty-stricken population in northeast China. *Cancer Manag Res* (2019) 11:5101–12. doi: 10.2147/CMAR.S205867