Asia-Pacific Journal of Oncology Nursing 9 (2022) 100101

Contents lists available at ScienceDirect

Asia-Pacific Journal of Oncology Nursing

journal homepage: www.apjon.org

Original Article

Comparison of machine learning models for predicting the risk of breast cancer-related lymphedema in Chinese women

Xiumei Wu [a,1], Qiongyao Guan [b,1], Andy S.K. Cheng [c], Changhe Guan [d], Yan Su [b], Jingchi Jiang [e], Yingchun Zeng [f,*], Linghui Zeng [f], Boran Wang [d,*]

[a] Department of Nursing, Integrated Hospital of Traditional Chinse Medicine (Southern Medical University Cancer Center), Southern Medical University, Guangzhou, China
[b] Department of Nursing, Yunnan Cancer Hospital, Kunming, China
[c] Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hong Kong SAR, China
[d] Department of Computer Sciences, Harbin Institute of Technology, Shenzhen, China
[e] Department of Computer Sciences, Harbin Institute of Technology, Harbin, China
[f] School of Medicine, Zhejiang University City College, Hangzhou, China

ARTICLE INFO

Keywords:
Breast cancer-related lymphedema
Machine learning
K-nearest neighbors
Support vector machine
Multilayer perceptron
Prediction

ABSTRACT

Objective: Predictive models for the occurrence of cancer symptoms by using machine learning (ML) algorithms could be used to aid clinical decision-making in order to enhance the quality of cancer care. This study aimed to develop and validate a selection of classification models that used ML algorithms to predict the occurrence of breast cancer-related lymphedema (BCRL) among Chinese women.

Methods: This was a retrospective cohort study of consecutive cases that had been diagnosed with breast cancer, stages I-IV. Forty-eight variables were grouped into five feature sets. Five classification models with ML algorithms were developed, and the models' performance and the variables' relative importance were assessed accordingly.

Results: Of 370 eligible female participants, 91 had BCRL (24.6%). The mean age of this study sample was 49.89 (SD = 7.45). All participants had had breast cancer surgery, and more than half of them had had a modified radical mastectomy (n = 206, 55.5%). The mean follow-up time after breast cancer surgery was 28.73 months (SD = 11.71). Most of the tumors were either stage I (n = 49, 31.2%) or stage II (n = 252, 68.1%). More than half of the sample had had postoperative chemotherapy (n = 227, 61.4%). Overall, the logistic regression model achieved the best performance in terms of accuracy (91.6%), precision (82.1%), and recall (91.4%) for BCRL. Although this study included 48 predicting variables, we found that the five models required only 22 variables to achieve predictive performance. The most important variable was the number of positive lymph nodes, followed in descending order by the BCRL occurring on the same side as the surgery, a history of sentinel lymph node biopsy, a dietary preference for meat and fried food, and an exercise frequency of less than three times per week. These factors were the most influential predictors for enhancing the ML models' performance.

Conclusions: This study found that in the ML training dataset, the multilayer perceptron model and the logistic regression model were the best discrimination models for predicting the outcome of BCRL, and the *k*-nearest neighbors and support vector machine models demonstrated good calibration performance in the ML validation dataset. Future research will need to use large-sample datasets to establish a more robust ML model for predicting BCRL deeply and reliably.

Introduction

Breast cancer (BC) is the most common cancer in the world for women.[1] In China, BC is the second most common type of cancer overall, after lung cancer.[2] Due to the advancements in BC treatment in recent decades, the five-year survival rate of BC has improved significantly and is approximately 90%.[3,4] In China, however, the five-year relative survival rate of BC patients varies from 96.5% (stage I) to 74.8% (stage III).[5] Women who have been treated for BC may experience numerous health challenges throughout their survivorship, one of which is BC-related

* Corresponding authors.
E-mail addresses: chloezengyc@hotmail.co.uk (Y. Zeng), wangboran@hit.edu.cn (B. Wang).
[1] These authors have contributed equally to this work as joint first author.

https://doi.org/10.1016/j.apjon.2022.100101
Received 30 January 2022; Accepted 30 May 2022

lymphedema (BCRL).[6]

The prevalence of BCRL varies from 5% to 77%.[6–9] The highest prevalence rate occurs within 3 years after the BC surgery, and the arm swelling rate increases almost 1% per year.[9–11] Breast cancer-related lymphedema can be a distressing side effect of BC and includes symptoms of swelling and progressive fibrotic skin changes, thus resulting in decreased quality of life (QOL) for the patient. In addition, BCRL is associated with recurrent soft-tissue infections, thus contributing to health care costs that might be lessened with prospective surveillance and early intervention.[10,12] In fact, with appropriate early intervention, early-stage lymphedema can be reversed, and early intervention can reduce the incidence or decrease the severity of its occurrence.[10,13] Hence, early prediction and diagnosis of BCRL may prevent its progression and reduce its negative effects on patients' QOL.[9]

Early prediction and detection of cancer-related symptoms such as BCRL are essential for improving the health outcomes of cancer patients.[14–16] Traditional statistical methods, such as multiple linear regression tests, have been used to identify predictors of health-related outcomes, but those statistical methods require a clear hypothesis and a rigorous research design, and their calculations of parameters can be modified only by redesigning the respective studies.[17] In recent years, many advanced technologies, such as machine learning (ML), have emerged as valuable approaches for expressing parameters in cancer treatments and outcomes.[17,18]

Machine learning refers to a collection of algorithmic techniques for data representation and analysis that have been widely applied to cancer care research.[19] The approaches in ML tend to be more suitable than traditional statistical methods for problems involving numerous potential predictors.[20] Additionally, the application of ML algorithms could increase prediction accuracy, because ML models tend to be nonparametric and able to learn complex interactions among predictors.[21]Whereas ML modeling studies for predicting the occurrence of BCRL in countries such as the United States have been done,[22] similar research has not been conducted in China. Therefore, this study aimed to develop and validate a selection of classification models that used ML algorithms to predict the occurrence of BCRL among Chinese women in Western China.

Methods

This was a retrospective cohort study of consecutive cases diagnosed with stages I-IV BC during the period 2017–2020. Ethical approval was obtained from Yunnan Cancer Hospital (Approval No. KYLX202106). This study was conducted by telephone calls, so only oral informed consent was obtained from each participant.

Study sample

All adult female patients at the Yunnan Cancer Hospital who had undergone a radical mastectomy for BC during the period 2017–2020 were identified as eligible cases. The presence of BCRL was determined by measuring the arm circumferences of the patients, using a flexible tape measure at four locations on each arm: (1) at 4 cm proximal to the wrist, (2) at 15 cm proximal to the first measure, (3) at 4 cm proximal to the olecranon, and (4) at 15 cm proximal to the third measure. A change of more than 2 cm in the absolute circumference at any of these four measurements was taken to be lymphedema.[23] Measuring timepoints for examining changes of arm circumstance were conducted at pre-surgery and post-surgery at time of discharging from hospital, or at pre-chemotherapy and at the end of chemotherapy. BCRL status were retrieved from the medical database and verified by arm lymphedema telephone questionnaire. Patients for whom clinicopathologic information was lacking regarding their BC diagnosis and treatment were excluded after the nurse had collected the data. Patients who had cognitive and communication impairments and were unable to communicate via phone were also excluded.

Features

The study used a data collection sheet with five parts (see Table 1 for a complete list): (1) Sociodemographic information, including age, race, and education. (2) Clinical data, including body mass index (BMI), history of chronic disease, and surgical history. (3) Tumor characteristics, including pathological type, tumor stage, and tumor location. (4) Treatment information, including type of surgery, type of surgical incision, and grade of axillary lymph node dissection (ALND). (5) Behavior-related information was collected using questionnaires that we self-designed with a set of structured questions: (i) Did you exercise pre-operatively: yes or no? If yes, please answer the following questions: (a) Identify which types of exercise you did: walking; jogging; yoga, or others; (b) Identify the exercise duration: 30 min per time or less; and (c) Identify the exercise frequency: 3 times per week, or less often. (ii) Did you take any post-operative exercise: yes or no? If yes, please identify the types of exercise you did: BC-related rehabilitation activity, walking, jogging, yoga, and/or others. (iii) If yes for post-op exercise, please also (a) identify the exercise duration (less than 30 min per session; or longer than 30 min), and (b) state your exercise frequency (fewer than 3 times per week, or more often). (iv) Did you perform physical labor: (a) pre-operatively: yes or no? (b) Post-operatively: yes or no? (v) Identify your daily dietary preferences: meat, milk, soy milk, fried food, dessert, vegetables, fruits. (vi) Did you receive a BCRL-related health education program given by nurses: yes or no?

Data collection

Data were collected by a combination of retrospectively extracting medical records and by telephone interviews, as complementary methods. The first four parts of data collection were the outcome measures of sociodemographic information, clinical data, tumor characteristics, and treatment information, and were retrieved from medical records. The fifth and final part of data collection used structured questionnaires that were administered by a trained research nurse via telephone interviews to collect behavior-related information. These five parts of data collection measures yielded a total of 48 variables/features (see Table 1 for details). According to the recommended five to 10 samples-per-feature ratios for ML, we estimated that this study would need from 240 to 480 subjects. The study ultimately had 370 subjects, which was adequate for exploring ML and avoiding overfitting in training an ML classifier with 48 features.[24]

Table 1
The study's variables and their corresponding categories utilized in predicting BCRL.

Category	Variables
Demographics	Age at diagnosis; race; marital status; education; area of residence (urban or rural); job type; medical insurance; smoking status; alcohol use
Clinical data	BMI; Handedness; pregnancy history; chronic disease history; surgical history; postoperative complications; blood pressure; venous blood draw; intravenous injection; lymph nodes-positive; duration of postoperative follow-up; ALND levels; SLNB status; Total number of SLNB; HER-2 detection; Ki-67 detection; hormone receptor type; metastatic lymph nodes; axillary/supraclavicular lymph nodes; BCRL occurring on the as the same side as the surgery side
Tumor characteristics	Tumor pathological type; breast quadrant of the tumor; tumor stage
Treatment types	Preoperative neoadjuvant chemotherapy; postoperative chemotherapy; postoperative radiotherapy; surgical incision; types of surgery; cycles of chemotherapy; location of radiotherapy; endocrine therapy
Behavior-related information	BCRL knowledge level; physical labor status; exercise frequency; exercise types; duration of exercise; dietary preferences; given a BCRL health education program

ALND, axillary lymph node dissection; BCRL, breast cancer-related lymphedema; SLNB, sentinel lymph node biopsy; BMI, body mass index.

Data analysis

The statistical analyses of this study were performed using the R Statistical Package (R Foundation), Python3.8.1, and PyTorch1.10.1, and we used the analytic schema shown in Fig. 1. Patients' sociodemographic information, clinical data, tumor characteristics, and treatment types were entered into various ML models to examine the lymphedema outcomes. Before constructing the ML models, the input data were randomly segregated into training and test sets. Approximately 80% of the data were used for training the prediction models, and nearly 20% were used in the test dataset for verification. This study divided our data into training and test sets for cross-validating the results later. In the model training phase, we divided the 300-sample data randomly into training and test datasets. Training sets were used to train the model, and test sets verified the model. Data were selected randomly to improve the generalizability of the training model. Using the principle of splitting the data in an ML field, the data could be divided with ratios of 9 (training set):1 (validation set), 8:2, 7:3, or 6:4. However, the training data set of this study would have been too small if we had adopted 6:4 or 7:3, because they could lead to an obvious under-fitting or over-fitting. On the other hand, if 9:1 had been adopted, the validity of the model would have been difficult to evaluate because of the small number of validation sets. Hence, this study adopted a division ratio of 8:2 for splitting our data into training and validation sets.

In order to address the study objectives, we used standard evaluation indices to evaluate our ML models, as suggested by previous research,[25] including the accuracy, precision score (positive predict rate), recall score (sensitivity), AUC (area under the receiver operating characteristic curve), F1 score (the harmonic mean of precision and recall scores), and the precision–recall curve for the training dataset. We used the calibration plot (i.e., a plot showing whether the risk prediction of BCRL was accurate) to examine the performance of the validation dataset. A threshold of 0.5 indicated equal weighting of false-positive and false-negative errors for all ML models.[26] Furthermore, this study computed the relative importance of each feature (i.e., the predicting variables) included in the ML models. Feature importance was obtained with these ML models on the basis of their features during ML training. Feature importance demonstrated how much the prediction changed as the variables' values varied, with a higher feature importance indicating a variable's greater importance in predicting the risk for BCRL.

Results

Sample characteristics

Of the study's 370 eligible participants, 91 had BCRL (24.6%). The mean age of this study sample was 49.89 years (SD = 7.45), and most of the women were of Han ethnicity ($n = 300$, 82.2%). All of the women had had BC surgery, and more than half of them had had a modified radical mastectomy ($n = 206$, 55.5%). The mean follow-up time after their BC surgery was 28.73 months (SD = 11.71). Most of the tumors were either stage I ($n = 49$, 31.2%) or stage II ($n = 252$, 68.1%). More than half of the sample had had postoperative chemotherapy ($n = 227$, 61.4%). Most of the women had a habit of regularly exercising more often than three times per week ($n = 334$, 90.3%), mainly in the form of walking or jogging ($n = 307$, 88.4%). A majority of the patients had an axillary lymph node dissection (ALND) severity level of I or II ($n = 273$, 73.8%). Very few had experienced BCRL health education programs provided by healthcare professionals ($n = 42$, 11.4%). Detailed characteristics of the participants are shown Table 2.

The ML models' prediction performance

There were 300 subjects included in the testing set. This study constructed five ML models for predicting BCRL: a naïve Bayes model, a *k*-nearest neighbor (KNN) model, a support vector machine (SVM) model, a logistic regression model, and a multilayer perceptron (MLP) model. The performance characteristics of the five ML models for predicting BCRL are summarized in Table 3. To assess the ML models' performance in outcome prediction, we divided the population into two categories: those with BCRL and those without BCRL. Overall, the logistic regression model achieved the best performance, with 91.6% accuracy, 82.1% precision, and 91.4% recall for BCRL (Table 3). The logistic regression model also had the highest F1 score (0.865), which was the harmonic mean of precision and recall, and the SVM model followed.

Among these five ML models, the MLP model had the largest area under the receiver operating characteristic curve, at 0.975 (Table 3 & Fig. 2). An MLP model is a classifier of neural networks that can be improved by increasing the number of neurons in the hidden layer. Different training and learning rules can be applied for training neural networks to enhance the performance of ML models. Hence, MLP models ultimately may be superior in their ability to predict whether an individual is likely to develop BCRL following surgery.

The performances of the five ML models in the validation set with 70 subjects are presented by the calibration plot in Fig. 3. Among the five ML algorithms, the KNN classifier was closest to the perfect calibration line and therefore had the best calibration performance.

The relative importance of the predicting variables (features)

Although this study included 48 predicting variables, we found that for all of the models, only 22 of the variables were needed for prediction performance (Fig. 4). Fig. 4 presents the relative importance of each of

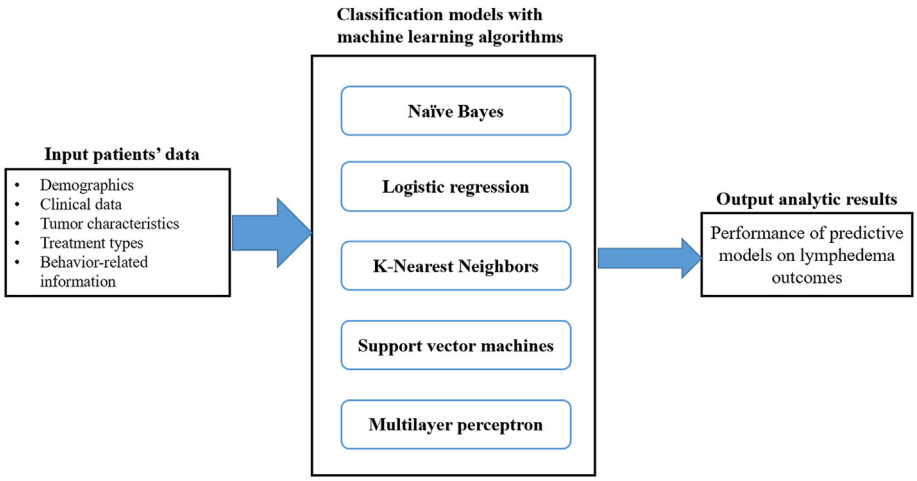

Fig. 1. Analytic schema for predictive model by the study's five classification models with machine-learning algorithms.

Table 2
Participants' characteristics.

Variables	*n* (%)	Variables	*n* (%)
Demographics		Tumor characteristics	
Age at diagnosis (M, SD)	49.89 (7.45)	Stage I	49 (13.2)
Race (Han)	300 (82.2)	Stage II	252 (68.1)
Marital status (married)	346 (93.5)	Stage III	43 (11.6)
Education (primary school)	270 (73.0)	Stage IV	26 (7.1)
Area of residence-urban	217 (58.6)	Breast quadrant of the tumor	
Area of residence-rural	153 (41.4)	Upper-outer	321 (86.8)
Follow-up, in months (M, SD)	28.73 (11.7)	Lower-outer	25 (6.7)
Job type (physical labor)	78 (21.1)	Upper-inner	19 (5.1)
Medical insurance (yes)	359 (97.0)	Lower-inner	5 (1.4)
Smoking status (yes)	15 (4.0)	**Treatment types**	
Alcohol use (yes)	14 (3.8)	Surgical side: Left	160 (43.2)
Clinical variables		Surgical side: Right	194 (52.4)
HER2 (positive)	52 (14.1)	Both sides	16 (4.4)
ER (positive)	11 (3.0)	Preoperative neoadjuvant chemotherapy (yes)	79 (21.4)
PR (positive)	43 (11.6)	Postoperative chemotherapy (yes)	227 (61.4)
ER & PR (positive)	265 (71.6)	Postoperative radiotherapy (yes)	119 (32.2)
History of pregnancy (yes)	355 (95.9)	Cycles of chemotherapy> 4	21 (5.7)
ER & PR (negative)	51 (13.8)	Surgical incision with curve	56 (15.1)
History of chronic disease (yes)	68 (18.4)		
History of surgery (yes)	370 (100)	Fusiform incision	238 (64.4)
Metastatic lymph nodes		Mixed incisions	20 (5.4)
< 10	203 (54.9)	Others	56 (15.1)
≥ 10	46 (12.4)	Surgery type (modified radical mastectomy)	206 (55.7)
BCRL side was the same as surgery side (yes)	329 (88.9)	Locations of radiotherapy within breast	32 (8.6)
Intravenous injection at BCRL side (yes)	12 (3.2)	Endocrine therapy (yes)	9 (2.4)
Blood pressure at BCRL side (yes)	28 (7.6)	**Behavior-related information**	
Lymph nodes-positive < 10	65 (17.6)	Preoperative physical activity (yes)	334 (90.3)
≥ 10	289 (78.1)	Postoperative physical activity (yes)	319 (86.2)
ALND levels		Exercise type: walking or jogging	287 (77.6)
BC stage I, II	273 (73.8)	Exercise duration > 30 min per time	301 (81.4)
BC stage III	97 (26.2)	Exercise frequency > 3 times per week	254 (68.6)
SLNB status	54 (14.6)	BCRL knowledge levels	
Total number of SLNBs ≥ 10	133 (35.9)	Low	103 (27.8)
Axillary supraclavicular lymph nodes < 10	298 (80.5)	Moderate	147 (39.7)
Ki67-positive	367 (99.2)	High	120 (32.4)
BMI (M, SD)	23.35 (3.55)	Dietary preference for meat and fried food	193 (52.2)
Postoperative complications (yes)	89 (24.1)	Received a BCRL health education program (yes)	42 (11.4)

BC, breast cancer; BCRL, breast cancer-related lymphedema; ER, estrogen receptor; HER-2, human epidermal growth factor receptor-2; PR, progesterone receptor; ALND, axillary lymph node dissection; BCRL, breast cancer-related lymphedema; SLNB, sentinel lymph node biopsy; BMI, body mass index.

Table 3
Performance results for the machine learning models in the validation set.

Model	AUC	Accuracy	Precision	Recall	F1
Naïve Bayes	0.893	0.883	0.818	0.771	0.794
Logistic regression	0.965	0.916	0.821	0.914	0.865
K-nearest neighbor	0.824	0.783	0.588	0.857	0.698
Support vector machine	0.948	0.892	0.775	0.886	0.827
Multilayer perceptron	0.975	0.875	0.708	0.971	0.819

AUC, area under the receiver operating characteristic curve.

those 22 features, in descending order. The most important feature was the number of positive lymph nodes, followed in descending order by the BCRL occurring on the same side as the surgery side, the presence of sentinel lymph node biopsies (SLNB), a dietary preference for meat and fried food, an exercise frequency of less than 3 times per week, and the number of months after surgery. These factors were the most influential predictive factors for enhancing the ML models' performance.

Discussion

The primary aim of this study was to apply machine learning algorithms to develop models that could successfully predict the breast cancer outcome of breast cancer-related lymphedema. This study demonstrated that five ML models could accurately predict the occurrence of BCRL. The

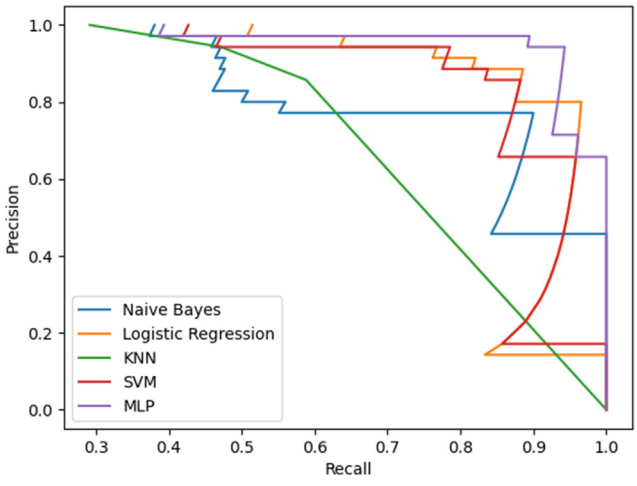

Fig. 2. Precision–recall curves of the five models in the validation set. KNN, *k*-nearest neighbors; SVM, support vector machine; MLP, multilayer perceptron

performance of discrimination (as measured by AUC) ranged from 0.824 to 0.975, with good to excellent accuracy (ranging from 0.783 to 0.916), and with good to excellent sensitivity (varying from 0.771 to 0.971).

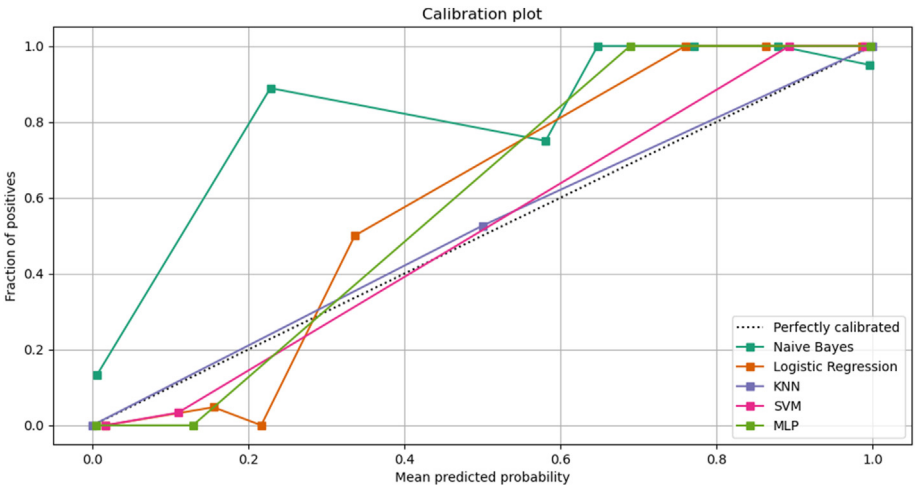

Fig. 3. Calibration plots.
KNN, *k*-nearest neighbors; SVM, support vector machine; MLP, multilayer perceptron.

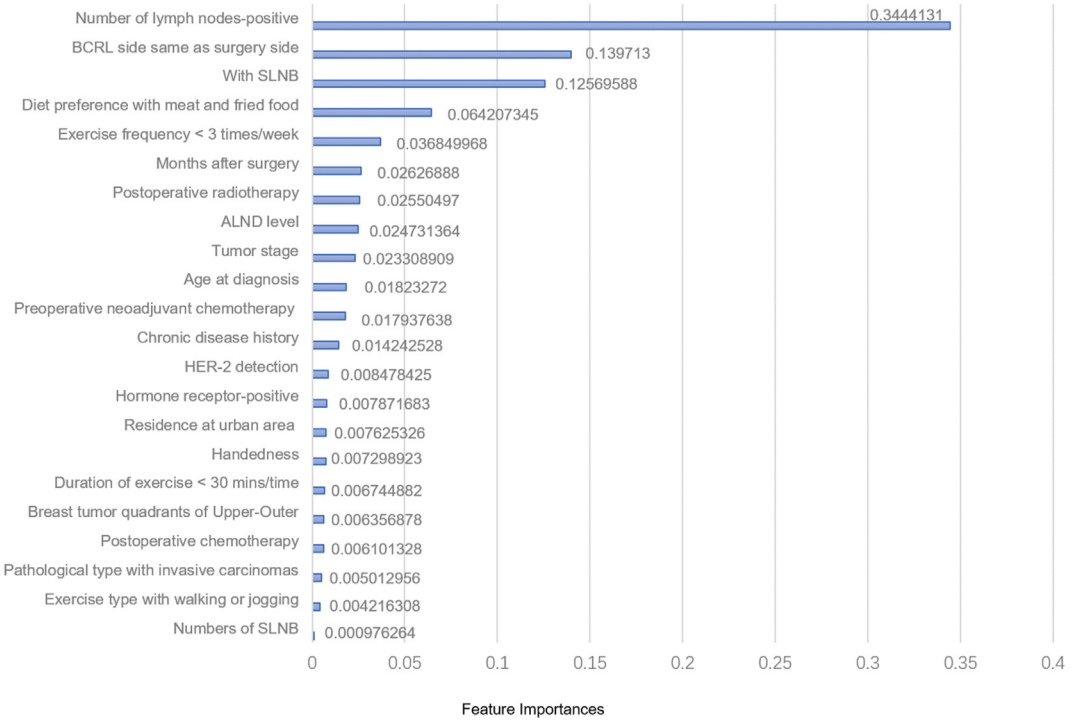

Fig. 4. The relative importance of each feature (variable), in descending order.
ALND, axillary lymph node dissection; HER-2, human epidermal growth factor receptor-2; SLNB, sentinel lymph node biopsy.

Among the five ML algorithms, the logistic regression model had the best performance, considering the trade-off between precision and sensitivity, with the highest F1 score of 0.865. In addition, we found that the KNN model had the best performance in terms of calibration.

The F1 score and the calibration value are both essential indicators of ML model performance.[27] The implications of one model performing best indicate that a well-trained ML classifier can offer more accurate predictions than other traditional approaches can for the occurrence of BCRL among patients after breast cancer.[22] If an increasing number of well-trained ML models provide highly accurate predictions of cancer prognosis, disease recurrence, or occurrence of symptoms, it is likely that the use of an ML classifier will become increasingly commonplace in many clinical settings.[28] Because machine learning belongs to data-driven technologies and is able to construct algorithms, it can continuously improve predictions and generate new knowledge for cancer symptom management.[22]

This study adopted five models—a naïve Bayes model, a KNN model, an SVM model, a logistic regression model, and an MLP model—for predicting the occurrence of BCRL. These five models, using ML for analysis, were based on the characteristics of the data, the sample size, or the sample distribution. The data features in this study were primarily numerical, the number of features was relatively varied, the sample distribution of each feature was highly differentiated, and the size of the data was relatively small. Consequently, these five models were selected because they were well-suited to the characteristics of the data. This study's prediction outcomes were binary classification problems—the outcome was either with or without BCRL—so that the logistic regression algorithm was very suitable for this study. In addition, the data sets for

this study were relatively small, and the Naïve Bayes model is generally applied for small data sets. On the other hand, the KNN approach is highly suitable for numerical data and is insensitive to outliers, which are common in medical data sets. In contrast, SVM's classification effect is better than those of the classifiers with the Naïve Bayes and KNN models, whereas an SVM model can effectively process high-dimensional feature data and establish nonlinear relationships between features. Finally, the MLP model approach incorporates the advantages of those ML classifiers and is highly suitable for the characteristics of the data set in this study. Notably, although a decision tree-based classifier is frequently used in disease outcome prediction in cancer care,[29] pruning of the decision tree is required for data sets with extremely high feature dimensions. Hence, this study omitted a decision tree-based classifier.

The most important finding of this study was that the ML prediction models identified several significant predictors of BCRL, the most important of which were the total number of positive lymph nodes, the BCRL being on the same side as the surgery side, a history with SLNB, a dietary preference for meat and fried food, and a low exercise frequency. Our findings were far different from the previous nomograms for predicting the risk of arm lymphedema after BC surgery developed by Bevilacqua et al,[15] who reported that patients' age, weight, height, level of axillary dissection, and radiotherapy field were the most important risk factors for arm lymphedema after breast cancer surgery. However, our study was consistent with Bevilacqua et al.'s results[15] in our finding that the number of months after BC surgery was an important predictor of BCRL.

Whereas the application of ML models for predicting BCRL remains in the early stages, further modeling studies could help clinicians to closely monitor the patients at risk of BCRL and to provide them with early referrals to cancer survivorship care. In particular, additional BCRL-related health education is needed, because we found that very few nurses are providing BCRL-related health education for patients.

This study had several limitations. First, the study was conducted only in a tumor hospital, so the generalizability of the results may be limited by the single-center setting. Second, some of the data were extracted retrospectively from medical records, so they cannot be guaranteed to contain all possible risk factors for BCRL. Third, the study's self-report measures may result in biased estimates, although those presumably would have been distributed equally among all participants. Finally, this study used a small data sample. Future research will require a relatively large dataset to validate and enhance the generalizability of the prediction models established in this study. Nevertheless, our findings can lay the groundwork for future studies using ML for BCRL prediction in oncology research.

Despite this study's inherent limitations, it established five novel ML models, and each achieved good discrimination, over and above that of nomograms for predicting arm lymphedema after BC surgery.[15] Early and accurate prediction of BCRL could help patients and clinicians improve the quality of cancer care and reduce the occurrence of BCRL, thus promoting the quality of life of breast cancer survivors. Of course, because machine learning belongs to the artificial intelligence framework and is increasingly used in different aspects of cancer care,[30,31] the ML models established in this study will need to be further validated for their ability to offer valuable predictive information about the risk of BCRL and to guide clinical decision making.

Conclusions

This study constructed and evaluated the performance of five machine learning models in predicting BCRL among Chinese women. Our findings revealed that in the ML training dataset, the MLP and logistic regression models were the best discrimination models for predicting the outcome of BCRL, and the KNN and SVM models demonstrated good calibration performance in the ML validation dataset. Future research will need to use large-sample datasets to establish a more robust ML model for predicting BCRL deeply and reliably. In addition, we found that a relatively high number of positive lymph nodes, BCRL on the same side as breast surgery, a history of sentinel lymph node biopsy, certain dietary preferences, and low exercise frequency were the most important predictive factors for BCRL. The study's findings can enable oncology nurses to precisely identify the patients who are most likely to have lymphedema following BC surgery and whom the nurses should then target for adequate health education at the time of discharge and/or during follow-up care.

Funding

This study was supported by a grant from the National Natural Science Foundation of China (Grant No. 72004039).

Declaration of competing interest

None declared.

Author contributions

Study design: QG, ASKC, YS, YZ; Data acquisition: YS, QG; Formal analysis: CG, XW, JJ, BW, YZ; Supervision: BW, YZ; Writing – original draft: XW, QG, BW, LZ, YZ; Writing—review & editing: All authors.

Ethics statement

This study was approved by the Institutional Review Board of Yunnan Cancer Hospital (Approval No. KYLX202106).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apjon.2022.100101.

References

1. Sung H, Ferlay J, Siegel RL, 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(3):209–249.
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(7):783–791.
3. Brunelle CL, Roberts SA, Horick NK, et al. Integrating symptoms into the diagnostic criteria for breast cancer-related lymphedema: applying results from a prospective surveillance program. *Phys Ther*. 2020;100(12):2186–2197.
4. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA A Cancer J Clin*. 2022;72(1):7–33.
5. Zuo T, Zeng H, Li H, et al. The influence of stage at diagnosis and molecular subtype on breast cancer patient survival: a hospital-based multi-center study. *Chin J Cancer*. 2017;36(1):84. https://doi.org/10.1186/s40880-017-0250-3.
6. Jørgensen MG, Toyserkani NM, Hansen FG, Bygum A, Sørensen JA. The impact of lymphedema on health-related quality of life up to 10 years after breast cancer treatment. *NPJ Breast Cancer*. 2021;7(1):70. https://doi.org/10.1038/s41523-021-00276-y.
7. Zhou K, Wang D, He X, et al. Effectiveness of a multimodal standard nursing program on health-related quality of life in Chinese mainland female patients with breast cancer: protocol for a single-blind cluster randomized controlled trial. *BMC Cancer*. 2016;16(1):698. https://doi.org/10.1186/s12885-016-2726-y.
8. Wang X, Li H, Yang Y, Su D, Zhang T. Current status and content of clinical practice guidelines related to lymphedema prevention behavior after breast cancer surgery. *China General Med*. 2017;20(6):639–644 [in Chinese].
9. Havens LM, Brunelle CL, Gillespie TC, et al. Use of technology to facilitate a prospective surveillance program for breast cancer-related lymphedema at the Massachusetts General Hospital. *mHealth*. 2021;7:11. https://doi.org/10.21037/mhealth-19-218.
10. Penn IW, Chang YC, Chuang E, et al. Risk factors and prediction model for persistent breast-cancer-related lymphedema: a 5-year cohort study. *Support Care Cancer*. 2019;27(3):991–1000.
11. Petrek JA, Senie RT, Peters M, Rosen PP. Lymphedema in a cohort of breast carcinoma survivors 20 years after diagnosis. *Cancer*. 2001;92(6):1368–1377.
12. Armer JM, Ballman KV, McCall L, et al. Factors associated with lymphedema in women with node-positive breast cancer treated with neoadjuvant chemotherapy and axillary dissection. *JAMA Surg*. 2019;154(9):800–809. https://doi.org/10.1001/jamasurg.2019.1742.

13. Du Y, Guan X, Chang D, Du Y, Li Z. Clinical study on the risk prediction of postoperative lymphedema in patients with breast conserving surgery based on Bevilacqua's model, 01 *Chin J Gen Surg*. 2021;15:53–56 [in Chinese].
14. Zhuang Y, Pan Z, Li M, Liu Z, Zhang Y, Huang Q. The effect of evidence-based nursing program of progressive functional exercise of affected limbs on patients with breast cancer-related lymphoedema. *Am J Transl Res*. 2021 Apr 15;13(4):3626–3633.
15. Bevilacqua JL, Kattan MW, Changhong Y, et al. Nomograms for predicting the risk of arm lymphedema after axillary dissection in breast cancer. *Ann Surg Oncol*. 2012;19(8):2580–2589.
16. Jung C, Kim J, Seo YJ, et al. Who will continuously depend on compression to control persistent or progressive breast cancer-related lymphedema despite 2 Years of conservative care? *J Clin Med*. 2020;9(11):3640. https://doi.org/10.3390/jcm9113640.
17. Ting WC, Lu YA, Ho WC, Cheewakriangkrai C, Chang HR, Lin CL. Machine learning in prediction of second primary cancer and recurrence in colorectal cancer. *Int J Med Sci*. 2020;17(3):280–291.
18. Kesler SR, Rao A, Blayney DW, Oakley-Girvan IA, Karuturi M, Palesh O. Predicting long-term cognitive outcome following breast cancer with pre-treatment resting state fMRI and random forest machine learning. *Front Hum Neurosci*. 2017;11:555.
19. Deslauriers J, Ansado J, Marrelec G, Provost JS, Joanette Y. Increase of posterior connectivity in aging within the Ventral Attention Network: a functional connectivity analysis using independent component analysis. *Brain Res*. 2017;1657:288–296.
20. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinf*. 2007;8:25.
21. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255–260.
22. Fu MR, Wang Y, Li C, et al. Machine learning for detection of lymphedema among breast cancer survivors. *mHealth*. 2018;4:17. https://doi.org/10.21037/mhealth.2018.04.02.
23. Ugur S, Arıcı C, Yaprak M, et al. Risk factors of breast cancer-related lymphedema. *Lymphatic Res Biol*. 2013;11(2):72–75.
24. Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*. 2003;19:1484–1491.
25. Lo YT, Liao JC, Chen MH, Chang CM, Li CT. Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms. *BMC Med Inf Decis Making*. 2021 Oct 20;21(1):288.
26. James C, Ranson JM, Everson R, Llewellyn DJ. Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients. *JAMA Netw Open*. 2021;4(12), e2136553.
27. Maitra S. Prediction & Calibration Techniques to Optimize Performance of Machine Learning Models. https://towardsdatascience.com/calibration-techniques-of-machine-learning-models-d4f1a9c7a9cf. Accessed on March 3, 2022..
28. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inf*. 2007;2:59–77.
29. Suresh A, Udendhran R, Balamurgan M. Hybridized neural network and decision tree-based classifier for prognostic decision making in breast cancers. *Soft Comput*. 2020;24:7947–7953.
30. Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clin Epidemiol Glob Health*. 2019;7(3):293–299.
31. Cheng ASK, Guan Q, Su Y, Zhou P, Zeng Y. Integration of machine learning and blockchain technology in the healthcare field: a literature review and implications for cancer care. *Asia Pac J Oncol Nurs*. 2021;8(6):720–724.