Received: 15 March 2023 | Revised: 27 June 2023 | Accepted: 5 July 2023

DOI: 10.1111/cas.15917

Cancer Science WILEY

ORIGINAL ARTICLE

Machine learning approaches for predicting 5-year breast cancer survival: A multicenter study

Quynh Thi Nhu Nguyen[1] | Phung-Anh Nguyen[2,3,4] | Chun-Jung Wang[1] | Phan Thanh Phuc[4] | Ruo-Kai Lin[1] | Chin-Sheng Hung[5] | Nei-Hui Kuo[6] | Yu-Wen Cheng[1] | Shwu-Jiuan Lin[1] | Zong-You Hsieh[4] | Chi-Tsun Cheng[4] | Min-Huei Hsu[2,7] | Jason C. Hsu[2,3,4,8]

[1]School of Pharmacy, College of Pharmacy, Taipei Medical University, Taipei City, Taiwan

[2]Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei City, Taiwan

[3]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei City, Taiwan

[4]Research Center of Health Care Industry Data Science, College of Management, Taipei Medical University, Taipei City, Taiwan

[5]Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei City, Taiwan

[6]Oncology Center, Taipei Medical University Hospital, Taipei City, Taiwan

[7]Graduate Institute of Data Science, College of Management, Taipei Medical University, Taipei City, Taiwan

[8]International Ph.D. Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei City, Taiwan

Correspondence
Jason C. Hsu, International Ph.D. Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, 11F, 301, Yuantong Rd, Zhonghe District, New Taipei City 235, Taiwan.
Email: jasonhsu@tmu.edu.tw

Min-Huei Hsu, Graduate Institute of Data Science, College of Management; Taipei Medical University, 11F, 301, Yuantong Rd, Zhonghe District, New Taipei City 235, Taiwan.
Email: 701056@tmu.edu.tw

Funding information
Taipei Medical University, Grant/Award Number: TMU108-AE1-B42; Taiwan Ministry of Science and Technology, Grant/Award Number: MOST110-2321-B-038-003, MOST111-2321-B038-005 and NSTC112-2321-B-038-005

Abstract

The study used clinical data to develop a prediction model for breast cancer survival. Breast cancer prognostic factors were explored using machine learning techniques. We conducted a retrospective study using data from the Taipei Medical University Clinical Research Database, which contains electronic medical records from three affiliated hospitals in Taiwan. The study included female patients aged over 20 years who were diagnosed with primary breast cancer and had medical records in hospitals between January 1, 2009 and December 31, 2020. The data were divided into training and external testing datasets. Nine different machine learning algorithms were applied to develop the models. The performances of the algorithms were measured using the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score. A total of 3914 patients were included in the study. The highest AUC of 0.95 was observed with the artificial neural network model (accuracy, 0.90; sensitivity, 0.71; specificity, 0.73; PPV, 0.28; NPV, 0.94; and F1-score, 0.37). Other

Abbreviations: AI, Artificial intelligence; ANN, Artificial neural network; AUC, Area under the receiver operating characteristic curve; BMI, Body mass index; BRC, Breast cancer; CCI, Charlson Comorbidity Index; CeVD, Cerebrovascular disease; CHF, Congestive heart failure; COPD, Chronic obstructive pulmonary disease; DPP-4, Dipeptidyl peptidase 4; ER, Estrogen receptor; GBM, Gradient boosting machine; HER2, Human epidermal growth factor receptor 2; JAK-SAT, Janus kinase-signal transducer and activator of transcription; LDA, Linear discriminant analysis; LGBM, Light gradient boosting machine; LR, Logistic regression; MAPK, Mitogen-activated protein kinases; MI, Myocardial infarction; NPV, Negative predictive value; PI3K/Akt, Phosphoinositide 3-kinase/Protein kinase B; PPV, Positive predictive value; PR, Progesterone receptor; PUD, Peptic ulcer disease; PVD, Peripheral vascular disease; RF, Random forest; SHH, Shuang-Ho Hospital; TCR, Taiwan Cancer Registry; TMUCRD, Taipei Medical University Clinical Research Database; TMUH, Taipei Medical University Hospital; WFH, Wan-Fang Hospital; XGBoost, Extreme gradient boosting.

models showed relatively high AUC, ranging from 0.75 to 0.83. According to the optimal model results, cancer stage, tumor size, diagnosis age, surgery, and body mass index were the most critical factors for predicting breast cancer survival. The study successfully established accurate 5-year survival predictive models for breast cancer. Furthermore, the study found key factors that could affect breast cancer survival in Taiwanese women. Its results might be used as a reference for the clinical practice of breast cancer treatment.

KEYWORDS
breast cancer survival, machine learning, prediction models, real-world data, TMUCRD

1 | INTRODUCTION

Breast cancer (BRC) is the most common cancer and the leading cause of death for women with cancer globally.[1] In the United States, there were an estimated 287,850 new cases and 43,250 female breast cancer deaths in 2022.[2] The incidence and mortality rates vary across racial groups and regions worldwide.[3,4] Prognostic factors of breast cancer can be divided into three groups: patient characteristics, such as age[5]; cancer characteristics, which include tumor size and lymph node status[6]; and biomarkers, which are measured from tumor cells, such as HER2, and hormone receptor status.[7] A prognostic prediction tool can support physicians in deciding appropriate treatment plans, which could enhance treatment effectiveness or lessen the suffering of patients.

Epidemiological studies play an important role in identifying prognostic factors of breast cancer, giving physicians some information for decision-making. However, the findings from these studies are not appropriate for patient-level prediction, and traditional statistical approaches are limited in the number of independent variables that can be included in the model.[8] To address this problem, many tools have been developed to predict survival outcomes. Two famous online prediction tools for breast cancer are Predict and Adjuvant! Online.[9,10] These tools were developed and validated using data from the United Kingdom, the United States, France, and Netherlands.[11–13] Other external validations made in Asian populations have revealed conflicting results. Both models showed overoptimistic prediction in a young Southeast Asian group (age < 40 years).[14] Predict underestimated overall survival in Japanese patients over 65 years,[15] Adjuvant! Online showed less accurate results in the high-risk group of Taiwanese patients.[16] Most machine learning models focus on cancer characteristics such as lymph nodes, tumor size, and biomarkers.[17–19] To date, few models have considered the effects of comorbidities and long-term medications on breast cancer prognosis.

The general health of cancer patients can also impact survival rates. Breast cancer patients with moderate and severe comorbidities have a higher risk of death.[20,21] Laboratory studies suggested anti-cancer effects of long-term medications such as aspirin,[22–24] statins,[25] beta-blockers,[26] ACE inhibitors, and ARBs[27] on breast cancer. Routine blood tests can reflect the overall health of the patients and are often used by physicians when assessing cancer prognosis. In a univariate model by Zhu et al,[28] breast cancer patients with normal red blood cell count, hematocrit, and albumin had a lower risk of recurrence compared to patients with lower corresponding parameters.

In this study, we aimed to develop prediction models for breast cancer patients based on demographic information, cancer characteristics, and other factors such as chronic diseases, long-term drugs, and laboratory exams. We also explored important prognostic factors of breast cancer using machine learning techniques.

2 | METHODS

2.1 | Data source

This study obtained data from Taipei Medical University Research Database (TMUCRD) from January 1, 2008 to December 31, 2020. The database combines the comprehensive data from three medical centers (i.e., Taipei Medical University Hospital [TMUH], Wan-Fang Hospital [WFH], and Shuang-Ho Hospital [SHH]) in the North of Taiwan. It is linked to the Taiwan Cancer Registry (TCR) and Taiwan Death Registry (TDR) databases that were established in 1979 and managed by Taiwan's Health Promotion Administration, Ministry of Health and Welfare. Furthermore, the TMUCRD contains the electronic medical record data of more than four million people from 1998 to 2021, including structured and unstructured data. This study has been approved by the Joint Institute Review Board of Taipei Medical University, Taipei, Taiwan. The data were anonymized before further analysis.

2.2 | Study design and cohort selection

We conducted a retrospective study in which we identified all female patients diagnosed with primary breast cancer (International Classification of Disease for Oncology, third edition [ICD-O-3] codes C50) from January 1, 2009 to December 31, 2019 in the TCR

database. We excluded subjects who were younger than 20 years and those who did not have any medical history in the three hospitals. Finally, 3914 patients were included in the study (Figure 1).

2.3 | Outcome measurement

We defined the breast cancer diagnosis date as the index date, and the study's outcome was 5-year survival after the index date. Medical records were reviewed for in-hospital deaths, and the TDR[29] was referred to in order to confirm the death status from inside and outside hospitals. The data were censored on the outcome date, at loss to follow-up (e.g., terminated national health insurance), or at the end of the study on December 31, 2020.

2.4 | Features selection

We selected those features that may lead to the death of BRC patients based on the literature review and the clinicians' consultations to develop the prediction models. All features were collected from outpatients and inpatients datasets. The variables were as follows:

1. Demographic information included age, body mass index (BMI), smoking, drinking, and betel chewing.
2. Cancer conditions included tumor size, cancer stage, biomarkers (e.g., human epidermal growth factor receptor 2 [HER2], estrogen receptor [ER], and progesterone receptor [PR]), and cancer treatments (e.g., surgery, radiotherapy). We observed patients' cancer conditions for 1 month after the cancer diagnosis.
3. Comorbidities included cardiovascular problems (i.e., consisting of myocardial infarction [MI], congestive heart failure [CHF], peripheral vascular disease [PVD], cerebrovascular disease), chronic obstructive pulmonary disease (COPD), rheumatic disease, peptic ulcer disease (PUD), renal disease, liver disease, diabetes, hyperlipidemia, hypertension, dementia, the and Charlson Comorbidity Index (CCI) score. These conditions were considered when patients were diagnosed over two or more outpatient visits or at an admission over a year before the index date.
4. Long-term medications were considered with antiplatelets, statins, biguanides, coxibs, benzodiazepines, beta-blockers, calcium channel blockers, angiotensin II receptor blockers, sulfonylureas, and dipeptidyl peptidase 4 (DPP-4). The medication uses were measured when patients received those for more than 1 month (30 days) during 1 year (360 days) before the BRC diagnosis.
5. Laboratory tests included tests for creatinine, fasting glucose, white blood cells, red blood cell, and platelets. We selected the current laboratory test values 1 year before or 3 months after the index date.

2.5 | Prediction model development

Several algorithms were selected to develop prediction models that can be formulated as classification models (i.e., binary outcomes). Those algorithms included logistic regression (LR), linear discriminant analysis (LDA), light gradient boosting machine (LGBM), gradient boosting machine (GBM), random forest (RF), AdaBoost, extreme gradient boosting (XGBoost), voting ensemble, and artificial neural network (ANN). A brief introduction to their parameters' settings is provided in S1 of Appendix S1.

2.6 | Model training and testing

In this study, prediction models were developed based on nine algorithms. The training dataset included the patient data from TMUH and WFH. We used the stratified fivefold cross-validation method in the training set to assess the performance of different algorithms and the overall errors. In detail, the dataset was divided into five subsets; each was used repeatedly as the internal validation set. Afterward, we used the patient data from SHH as the external testing set to evaluate the models' generalization.

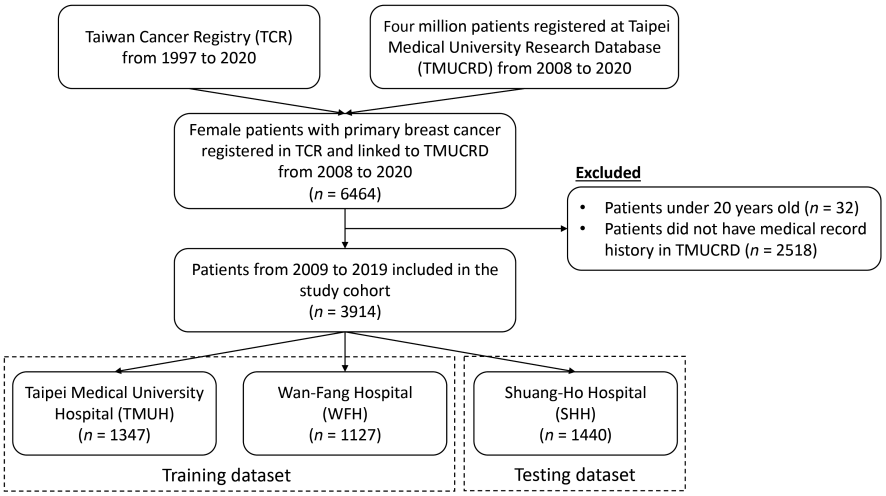

FIGURE 1 Cohort selection process.

2.7 | Model performance

The performances of the algorithms were measured using the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity (recall), specificity, positive predictive value (PPV, precision), negative predictive value (NPV), and F1-score. The best model was defined as the highest AUC by comparing various models based on the external testing set. We analyzed the feature's contribution (i.e., the feature's importance) to the best model using SHapley Additive exPlanations (SHAP) values.[30]

All the data processing was performed using the MSSQL server 2017, the machine learning algorithms were generated using Scikit-Learn library version 1.0.2, and the ANN model was developed with Tensor Flow version 2.9.0 in Python programing language version 3.9.[31]

3 | RESULTS

3.1 | Baseline characteristics of study cohorts

We identified 6464 eligible patients diagnosed with primary breast cancer and registered at TCR from 2008 to 2020. We excluded 32 patients younger than 20 years and 2518 patients with no medical history in TMUCRD at the index date. A total of 3914 patients were included in the study, in which 2474 patients were assigned to the training dataset, whereas 1440 patients were included in the testing dataset.

Table 1 shows the basic characteristics of the study cohort, including patients' demographic information, cancer conditions, comorbidities, current medications, and laboratory test results. The mean (standard deviation, SD) ages and BMI of cohort patients were 55.6 (12.4) and 24.2 (4.26), respectively. Most patients with early-stage breast cancer (i.e., stage I, 28.1% and stage II, 35.8%) and a high proportion received surgery (73.2%). The cohort of patients had comorbidities related to hypertension (18.3%), hyperlipidemia (15.7%), and cardiovascular problems (10.9%). The overall mean (SD) CCI score was 3.80 (1.88). Patients received benzodiazepine with the highest proportion (17%), followed by statin (9.4%), antiplatelets (8.8%), and angiotensin II receptor blockers (8.7%). The mortality rates for the training and testing cohort dataset were 7% and 10.2%, respectively. Detailed information is shown in Table S1 in Appendix S1. The associations between different features and the outcome at the patient baseline are shown in Table S2 in Appendix S1.

3.2 | The performances of different prediction models

Table 2 shows the performance of the survival prediction models. The highest AUC of 0.95 was observed with the ANN model (i.e., accuracy, 0.90; sensitivity, 0.71; specificity, 0.73, PPV, 0.28; NPV, 0.94; and F1-score, 0.37) compared to other models. Among the machine learning algorithms, the AUC of the voting ensemble model was observed as the highest, at 0.83 (i.e., accuracy, 0.68, sensitivity, 0.85; specificity, 0.66; and F1-score, 0.60), followed by the RF, and AdaBoost models with an AUC of 0.82. Figure 2 shows the receiver operator characteristic curves of various models. The precision-recall curve of different machine learning models is shown in Figure S1 in Appendix S1. Figure 3 shows the feature importance of the ANN model. The most important features were cancer stage, tumor size, age at diagnosis, BMI, and other biomarkers.

4 | DISCUSSION

In this study, ML models were developed using Taipei Medical University Clinical Research Database data to predict the 5-year survival of breast cancer patients. All models showed relatively high AUC, ranging from 0.75 (logistic regression) to 0.83 (voting classifier). We also used a deep learning technique to build a model (ANN), which showed the best performance overall (AUC, 0.95; accuracy, 0.90; sensitivity, 0.71; specificity, 0.73; PPV, 0.28; NPV, 0.94; and F1-score, 0.37). In addition, the relationship between features and prediction models' accuracy was also examined.

Machine learning techniques have been applied to molecular property prediction in drug development for a decade. Several studies used genomic data to predict the survival of breast cancer cell lines, which assisted the drug-response assessment in drug discovery and repositioning.[32,33] In contrast, machine learning and deep learning studies focus on clinical data and their applications to patient-level prediction for breast cancer are limited. Studies by Ganggayah et al.,[17] Xiao et al.,[18] and Huang et al.[19] using machine learning algorithms to predict the overall survival of breast cancer patients showed comparable performance to our research. Although RF was not the best among those algorithms, it performed well in all four studies. This finding indicates that RF is particularly suitable for prognosis prediction tasks, which can be explained by its ability to handle nonlinear data and reduced tendency to overfit.[34] In another work, Ganggayah's team[35] also developed one deep learning neural network (multilayer perceptron), which showed 88.2% accuracy in the testing set. Our deep learning model (ANN) obtained higher AUC and accuracy (0.95 and 0.90, respectively).

This study reinforced the findings from previous work. Tumor size and cancer stage were the two most important features of the prediction model. A study by Han et al. using data from breast cancer patients from the United States reported that tumor size and lymph node metastasis were significantly associated with overall survival.[36] These variables were used in almost all studies for survival analysis and showed a high correlation with the death of breast cancer patients.[17–19,35] Another strong predictor observed in our study was BMI. The association between obesity and breast cancer has long been a topic of interest to many researchers. Being overweight or obese not only increases the risk but also has an impact on breast cancer progression. Leptin, an adipokine produced by adipose tissue, activates multiple signaling pathways, including Janus

TABLE 1 Basic characteristics of the study cohort.

	Overall ($n = 3914$)	Training cohort ($n = 2474$)[a]	Testing cohort ($n = 1440$)[b]
5-year mortality, *N* (%)	321 (8.2)	174 (7.0)	147 (10.2)
Demographic information			
Age, mean (SD), yrs.	55.6 (12.4)	55.3 (12.7)	56.1 (11.9)
BMI, mean (SD), kg/m^2	24.2 (4.26)	24.0 (4.20)	24.6 (4.33)
Smoking, *N* (%)			
No	2683 (68.5)	1688 (68.2)	995 (69.1)
Yes	180 (4.6)	99 (4.0)	81 (5.6)
Unknown	1051 (26.9)	687 (27.8)	364 (25.3)
Drinking, *N* (%)			
No	2646 (67.6)	1647 (66.6)	999 (69.4)
Yes	177 (4.5)	132 (5.3)	45 (3.1)
Unknown	1091 (27.9)	695 (28.1)	396 (27.5)
Betel chewing, *N* (%)			
No	2877 (73.5)	1797 (72.6)	1080 (75.0)
Yes	4 (0.1)	3 (0.1)	1 (0.1)
Unknown	1033 (26.4)	674 (27.2)	359 (24.9)
Cancer condition			
Tumor size, mm			
Mean (SD)	24.7 (19.5)	24.2 (19.5)	25.6 (19.4)
Median [IQR]	20 [13–30]	20 [12–30]	21 [14–32]
Cancer stage, *N* (%)			
Stage = 0	674 (17.2)	537 (21.7)	137 (9.5)
Stage = 1	1098 (28.1)	765 (30.9)	333 (23.1)
Stage = 2	1402 (35.8)	867 (35.0)	535 (37.2)
Stage = 3	153 (3.9)	90 (3.6)	63 (4.4)
Stage = 4	169 (4.3)	82 (3.3)	87 (6.0)
Unknown	418 (10.7)	133 (5.4)	285 (19.8)
HER2, *N* (%)			
Negative	1967 (50.3)	1244 (50.3)	723 (50.2)
Positive	641 (16.4)	381 (15.4)	260 (18.1)
Unknown	1306 (33.4)	849 (34.3)	457 (31.7)
PR, *N* (%)			
Negative	781 (20.0)	500 (20.2)	281 (19.5)
Positive	2141 (54.7)	1365 (55.2)	776 (53.9)
Unknown	992 (25.3)	609 (24.6)	383 (26.6)
ER, *N* (%)			
Negative	558 (14.3)	328 (13.3)	230 (16.0)
Positive	2369 (60.5)	1540 (62.2)	829 (57.6)
Unknown	987 (25.2)	606 (24.5)	381 (26.5)
Radiation therapy, *N* (%)			
No	1348 (34.4)	996 (40.3)	352 (24.4)
Yes	1711 (43.7)	942 (38.1)	769 (53.4)
Unknown	855 (21.8)	536 (21.7)	319 (22.2)

(Continues)

TABLE 1 (Continued)

	Overall ($n = 3914$)	Training cohort ($n = 2474$)[a]	Testing cohort ($n = 1440$)[b]
Surgery, *N* (%)			
No	205 (5.2)	123 (5.0)	82 (5.7)
Yes	2866 (73.2)	1826 (73.8)	1040 (72.2)
Unknown	843 (21.5)	525 (21.2)	318 (22.1)
Comorbidity, *N* (%)			
Cardiovascular problems[c]	426 (10.9)	273 (11.0)	153 (10.6)
Dementia	178 (4.5)	125 (5.1)	53 (3.7)
COPD	350 (8.9)	270 (10.9)	80 (5.6)
Rheumatic disease	115 (2.9)	82 (3.3)	33 (2.3)
PUD	487 (12.4)	315 (12.7)	172 (11.9)
Renal disease	91 (2.3)	62 (2.5)	29 (2.0)
Liver disease	308 (7.9)	229 (9.3)	79 (5.5)
Diabetes	186 (4.8)	98 (4.0)	88 (6.1)
Hyperlipidemia	614 (15.7)	428 (17.3)	186 (12.9)
Hypertension	715 (18.3)	467 (18.9)	248 (17.2)
CCI score			
Mean (SD)	3.80 (1.88)	3.82 (1.96)	3.75 (1.75)
Median [IQR]	3.0 [2.0–5.0]	3.0 [2.0–5.0]	3.0 [2.0–5.0]
Medication (ATC code), *N* (%)			
Beta blocking agents (C07AB)	242 (6.2)	142 (5.7)	100 (6.9)
Calcium channel blockers (C08CA)	315 (8.0)	187 (7.6)	128 (8.9)
Angiotensin II receptor blockers (C09CA)	340 (8.7)	191 (7.7)	149 (10.3)
Biguanides (A10BA)	194 (5.0)	106 (4.3)	88 (6.1)
DPP-4 (A10BH)	96 (2.5)	52 (2.1)	44 (3.1)
Sulfonylureas (A10BB)	127 (3.2)	63 (2.5)	64 (4.4)
Statins (C10AA)	366 (9.4)	209 (8.4)	157 (10.9)
Antiplatelets (B01AC)	344 (8.8)	187 (7.6)	157 (10.9)
Coxibs (M01AH)	313 (8.0)	154 (6.2)	159 (11.0)
Benzodiazepines (N05BA)	667 (17.0)	341 (13.8)	326 (22.6)
Laboratory test, Mean (SD)			
Creatinine	0.84 (0.98)	0.80 (0.83)	0.92 (1.21)
WBC	7.41 (3.05)	7.02 (2.69)	8.23 (3.56)
RBC	4.33 (0.59)	4.36 (0.56)	4.29 (0.64)
Platelet (PLT)	250 (90.2)	247 (81.5)	256 (106)
Fasting glucose	117 (41.8)	114 (40.8)	123 (43.7)

Abbreviations: BMI, body mass index; COPD, Chronic obstructive pulmonary disease; CVD, cardiovascular; DPP-4, dipeptidyl peptidase-4; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; IQR, interquartile range; PLT, platelet; PR, progesterone receptor; PUD, peptic ulcer disease; RBC, red blood cell count; SD, standard deviation; WBC, white blood cell count; yrs., years.

[a]The training set included data from Taipei Medical University and Wan-Fang Hospital.

[b]The testing set included data from Shuang Ho Hospital.

[c]Cardiovascular problems consisted of myocardial infarction (MI), congestive heart failure (CHF), peripheral vascular disease (PVD), and cerebrovascular disease.

kinase-signal transducer and activator of transcription, mitogen-activated protein kinases, and phosphoinositide 3-kinase/protein kinase B. These pathways induce immigration and invasion of tumor cells, angiogenesis, and recruitment of immune cells.[37–39]

As our study focused on the overall deaths of breast cancer patients, we took into consideration not only breast cancer-specific factors but also general health-related factors. Another important feature of our model was CCI score, a tool used for over

TABLE 2 Performance of survival prediction models.

Model	Training AUC	Testing AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	F1-score
Logistic regression	0.80	0.75	0.79	0.59	0.81	0.26	0.95	0.44
Linear discriminant analysis	0.84	0.78	0.72	0.74	0.72	0.23	0.96	0.54
LGBM classifier	0.99	0.81	0.77	0.71	0.78	0.27	0.96	0.57
Gradient boosting classifier	0.94	0.81	0.72	0.77	0.72	0.24	0.97	0.55
XGB classifier	1.00	0.78	0.72	0.71	0.72	0.22	0.96	0.53
Random forest	0.87	0.82	0.78	0.71	0.78	0.27	0.96	0.57
Ada boost classifier	0.91	0.82	0.80	0.71	0.81	0.30	0.96	0.52
Voting classifier	0.92	0.83	0.68	0.85	0.66	0.22	0.98	0.60
ANN	0.98	0.95	0.90	0.71	0.73	0.28	0.94	0.37

Abbreviations: ANN, artificial neural network; AUC, area under the curve; LGBM, light gradient boosting machine; NPV, negative prediction value; PPV, positive prediction value; XGB, extreme gradient boosting.

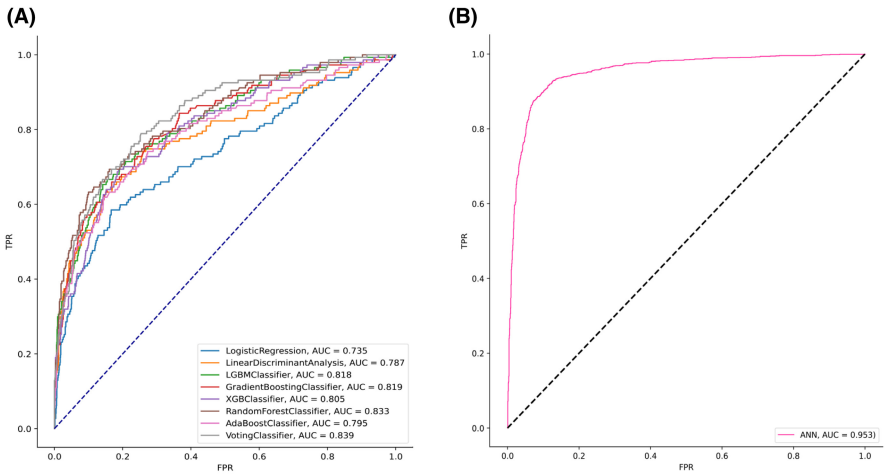

FIGURE 2 The performance of the prediction models in the testing dataset. (A) Receiver operator characteristic (ROC) curve of different machine learning models. (B) ROC curve of the artificial neural network model.

30 years by clinicians to assess the prognosis of various cancer types and other severe health conditions. Although several epidemiological studies have validated it,[40–43] this variable was not considered in previous machine learning studies that had a similar aim to ours,[17–19,35] as these studies mainly focused on tumor characteristics. Hypertension, a comorbidity not included in the CCI, was another variable that contributed to the models' performance. The prevalence of hypertension is high among breast cancer patients, especially in the older group.[44–46] Jung et al. found that hypertension was associated with a higher mortality risk in patients with metastatic breast cancer even when age and other covariates were adjusted.[47]

The present study acknowledges several limitations. First, the retrospective design of the study warrants caution in generalizing the findings, necessitating further research employing a prospective design to validate the models. Second, although data from multiple sites (TMUH and WFH for training and SHH for external testing) were utilized, it is important to note that all these hospitals are located in northern Taiwan, which might limit the representation of the entire Taiwanese population. To enhance the model's validity, future investigations will incorporate data from diverse regions of Taiwan and other Asian countries, including Korea, Japan, Singapore, Australia, and China. Third, the integration of laboratory and genomic data has the potential to enhance the performance of machine learning models. However, due to the unavailability of many of these data points, they were not included in this study. Fourth, unlike similar studies, this model did not encompass drug therapy. The focus was on patients newly diagnosed with breast cancer who were monitored over 1 month, during which time only a small subset of patients received drug therapy, while surgery and radiation therapy were predominantly administered at the onset of treatment. Finally, the limited sample size necessitated the development of models that provide probabilities for outcomes rather than risk levels. This limitation can be addressed in future studies as more extensive data are accumulated.

In the current study, we built machine learning models to analyze breast cancer patients' 5-year survival. The most important prognostic factors identified in this study were cancer stage, tumor size, diagnosis age, surgery, and BMI. The model using the ANN algorithm

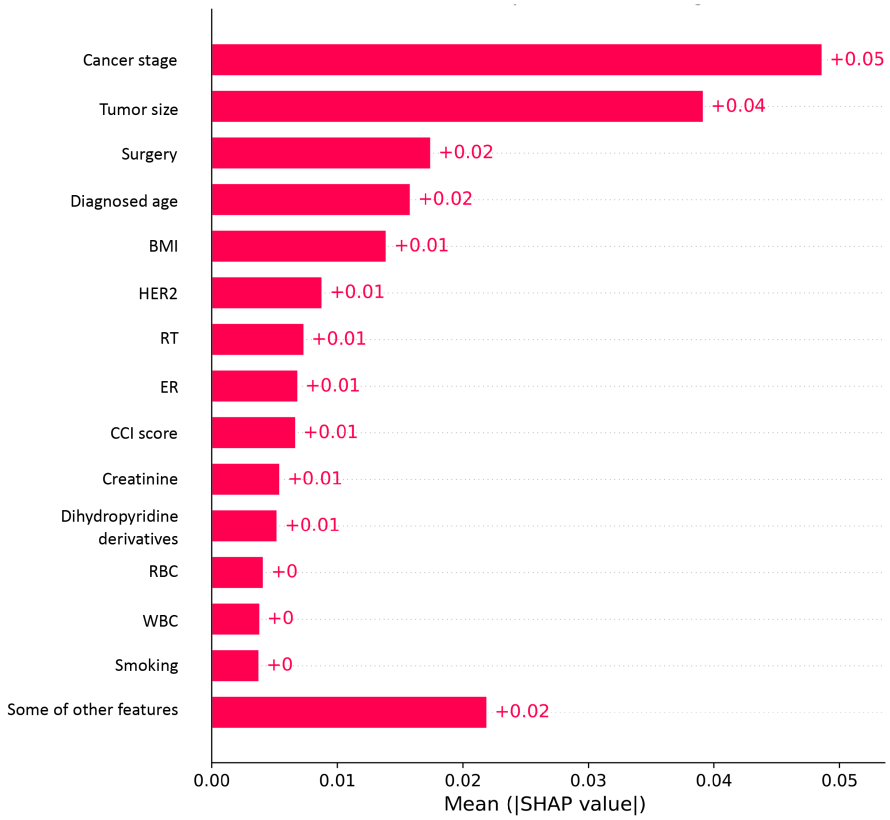

FIGURE 3 Feature importance of the artificial neural network prediction model.

yielded the best performance among all. Findings from this study identify directions for future work to improve the prediction model and to better understand the feasibility of applying this tool in clinical practice.

AUTHOR CONTRIBUTIONS

APAN, JCH, YTC, SCY, CCL, and YHY conceptualized and designed the study. APAN, YCL, TCH, YHF, PCL, PCH, HET, SCC, and WCC provided clinical research design suggestions. YTC, YCL, HCH, JSW, and CML collected data, performed the analyses, and drafted the manuscript. APAN and CYL reviewed all data and revised the manuscript critically for intellectual content. All authors approved the final version for submission.

ACKNOWLEDGMENTS

This work was supported by the Taiwan Ministry of Science and Technology grants (grant numbers MOST110-2321-B-038-003, MOST111-2321-B038-005, and NSTC112-2321-B-038-005) and the Taipei Medical University (grant number TMU108-AE1-B42).

CONFLICT OF INTEREST STATEMENT

All authors declare none.

DATA AVAILABLITY STATEMENT

The data source was hospital electronic medical records from three medical centers in Taiwan, including Taipei Medical University Hospital, Shuang-Ho Hospital, and Wan-Fang Hospital.

ETHICS STATEMENT

The study was conducted following the protocol approved by the Joint Institutional Review Boards of Taipei Medical University.
Informed consent: N/A.
Registry and the Registration No. of the study/trial: N/A.
Animal Studies: N/A.

PATIENT AND PUBLIC INVOLVEMENT

It was not appropriate or possible to involve patients or the public in the design, conduct, reporting, or dissemination plans of our research.

ORCID

Phung-Anh Nguyen https://orcid.org/0000-0002-7436-9041
Phan Thanh Phuc https://orcid.org/0000-0001-9132-1456
Shwu-Jiuan Lin https://orcid.org/0000-0002-6961-4890
Jason C. Hsu https://orcid.org/0000-0003-2997-2404

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;71:209-249.
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72:7-33.
3. Hu K, Ding P, Wu Y, Tian W, Pan T, Zhang S. Global patterns and trends in the breast cancer incidence and mortality according to sociodemographic indices: an observational study based on the global burden of diseases. *BMJ Open*. 2019;9:e028461.

4. Jatoi I, Sung H, Jemal A. The emergence of the racial disparity in U.S. breast-cancer mortality. *N Engl J Med*. 2022;386:2349-2352.
5. Høst H, Lund E. Age as a prognostic factor in breast cancer. *Cancer*. 1986;57:2217-2221.
6. Donegan WL. Tumor-related prognostic factors for breast cancer. *CA Cancer J Clin*. 1997;47:28-51.
7. Dawood S, Broglio K, Buzdar AU, Hortobagyi GN, Giordano SH. Prognosis of women with metastatic breast cancer by *HER2* status and trastuzumab treatment: an institutional-based review. *J Clin Oncol*. 2010;28:92-98.
8. Madakkatel I, Zhou A, McDonnell MD, Hyppönen E. Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. *Sci Rep*. 2021;11:11.
9. Ravdin PM, Siminoff LA, Davis GJ, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *J Clin Oncol*. 2001;19:980-991.
10. Wishart GC, Azzato EM, Greenberg DC, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res*. 2010;12:R1.
11. Hajage D, De Rycke Y, Bollet M, et al. External validation of adjuvant! Online breast cancer prognosis tool. Prioritising Recommendations for Improvement. *PloS One*. 2011;6:e27446.
12. Gray E, Marti J, Brewster DH, Wyatt JC, Hall PS. Independent validation of the PREDICT breast cancer prognosis prediction tool in 45,789 patients using Scottish cancer registry data. *Br J Cancer*. 2018;119:808-814.
13. Candido Dos Reis FJ, Wishart GC, Dicks EM, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Res*. 2017;19:19.
14. Bhoo-Pathy N, Yip C-H, Hartman M, et al. Adjuvant! Online is overoptimistic in predicting survival of Asian breast cancer patients. *Eur J Cancer*. 2012;48:982-989.
15. Zaguirre K, Kai M, Kubo M, et al. Validity of the prognostication tool PREDICT version 2.2 in Japanese breast cancer patients. *Cancer Med*. 2021;10:1605-1613.
16. Yao-Lung K, Dar-Ren C, Tsai-Wang C. Accuracy validation of adjuvant! Online in Taiwanese breast cancer patients–a 10-year analysis. *BMC Med Inform Decis Mak*. 2012;12:108.
17. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak*. 2019;19:48.
18. Xiao J, Mo M, Wang Z, et al. The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study. *JMIR Med Inform*. 2022;10:e33440.
19. Huang K, Zhang J, Yu Y, Lin Y, Song C. The impact of chemotherapy and survival prediction by machine learning in early elderly triple negative breast cancer (eTNBC): a population based study from the SEER database. *BMC Geriatr*. 2022;22:268.
20. Yancik R. Effect of age and comorbidity in postmenopausal breast cancer patients aged 55 years and older. *JAMA*. 2001;285:885-892.
21. Piccirillo JF. Prognostic importance of comorbidity in a hospital-based cancer registry. *JAMA*. 2004;291:2441.
22. Choi B-H, Chakraborty G, Baek K, Yoon HS. Aspirin-induced Bcl-2 translocation and its phosphorylation in the nucleus trigger apoptosis in breast cancer cells. *Exp Mol Med*. 2013;45:e47.
23. Tsujii M, DuBois RN. Alterations in cellular adhesion and apoptosis in epithelial cells overexpressing prostaglandin endoperoxide synthase 2. *Cell*. 1995;83:493-501.
24. Allen JE, Patel AS, Prabhu VV, et al. COX-2 drives metastatic breast cells from brain lesions into the cerebrospinal fluid and systemic circulation. *Cancer Res*. 2014;74:2385-2390.
25. Demierre M-F, Higgins PDR, Gruber SB, Hawk E, Lippman SM. Statins and cancer prevention. *Nat Rev Cancer*. 2005;5:930-942.
26. Powe DG, Voss MJ, Habashy HO, et al. Alpha- and beta-adrenergic receptor (AR) protein expression is associated with poor clinical outcome in breast cancer: an immunohistochemical study. *Breast Cancer Res Treat*. 2011;130:457-463.
27. Koh W-P, Yuan J-M, Van Den Berg D, Lee H-P, Yu MC. Polymorphisms in angiotensin II type 1 receptor and angiotensin I-converting enzyme genes and breast cancer risk among Chinese women in Singapore. *Carcinogenesis*. 2005;26:459-464.
28. Zhu Z, Li L, Ye Z, et al. Prognostic value of routine laboratory variables in prediction of breast cancer recurrence. *Sci Rep*. 2017;7:7.
29. Center TCR. *Report Data*. Taiwan Cancer Registration Center; 2022.
30. Ning Y, Ong MEH, Chakraborty B, et al. Shapley variable importance cloud for interpretable machine learning. *Patterns*. 2022;3:100452.
31. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
32. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med*. 2021;13:112.
33. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics*. 2021;22:214.
34. Touw WG, Bayjanov JR, Overmars L, et al. Data mining in the life sciences with random Forest: a walk in the park or lost in the jungle? *Brief Bioinform*. 2012;14:315-326.
35. Kalafi EY, Nor NAM, Taib NA, Ganggayah MD, Town C, Dhillon SK. Machine learning and deep learning approaches in breast cancer survival prediction using clinical data. *Folia Biol (Praha)*. 2019;65:212-220.
36. Han Y, Wang J, Sun Y, et al. Prognostic model and nomogram for estimating survival of small breast cancer: a SEER-based analysis. *Clin Breast Cancer*. 2021;21:e497-e505.
37. Alshaker H, Krell J, Frampton AE, et al. Leptin induces upregulation of sphingosine kinase 1 in oestrogen receptor-negative breast cancer via Src family kinase-mediated, janus kinase 2-independent pathway. *Breast Cancer Res*. 2014;16:426.
38. Haque I, Ghosh A, Acup S, et al. Leptin-induced ER-α-positive breast cancer cell viability and migration is mediated by suppressing CCN5-signaling via activating JAK/AKT/STAT-pathway. *BMC Cancer*. 2018;18:99.
39. Cao H, Huang Y, Wang L, et al. Leptin promotes migration and invasion of breast cancer cells by stimulating IL-8 production in M2 macrophages. *Oncotarget*. 2016;7:65441-65453.
40. Frenkel WJ, Jongerius EJ, Mandjes-Van Uitert MJ, Van Munster BC, De Rooij SE. Validation of the Charlson comorbidity index in acutely hospitalized elderly adults: a prospective cohort study. *J Am Geriatr Soc*. 2014;62:342-346.
41. Zhao L, Leung L-H, Wang J, et al. Association between Charlson comorbidity index score and outcome in patients with stage IIIB-IV non-small cell lung cancer. *BMC Pulm Med*. 2017;17:112.
42. Huang Y, Chen W, Haque W, et al. The impact of comorbidity on overall survival in elderly nasopharyngeal carcinoma patients: a National Cancer Data Base analysis. *Cancer Med*. 2018;7:1093-1101.
43. Moro-Sibilot D, Aubert A, Diab S, et al. Comorbidities and Charlson score in resected stage I nonsmall cell lung cancer. *Eur Respir J*. 2005;26:480-486.
44. Yancik R, Havlik RJ, Wesley MN, et al. Cancer and comorbidity in older patients: a descriptive profile. *Ann Epidemiol*. 1996;6:399-412.
45. Fleming ST, Rastogi A, Dmitrienko A, Johnson KD. A comprehensive prognostic index to predict survival based on multiple comorbidities: a focus on breast cancer. *Med Care*. 1999;37:601-614.

46. Gironés R, Torregrosa D, Díaz-Beveridge R. Comorbidity, disability and geriatric syndromes in elderly breast cancer survivors. Results of a single-center experience. *Crit Rev Oncol Hematol*. 2010;73:236-245.
47. Jung SY, Rosenzweig M, Linkov F, Brufsky A, Weissfeld JL, Sereika SM. Comorbidity as a mediator of survival disparity between younger and older women diagnosed with metastatic breast cancer. *Hypertension*. 2012;59:205-211.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Nguyen QTN, Nguyen P-A, Wang C-J, et al. Machine learning approaches for predicting 5-year breast cancer survival: A multicenter study. *Cancer Sci*. 2023;114:4063-4072. doi:10.1111/cas.15917