Machine Learning Models for Breast Lesions Based on Ultrasound Imaging Features: A Observational Study

Yao Tan
Peking University First Hospital
Ling Huo
Peking University Cancer Hospital: Beijing Cancer Hospital
Shu Wang
Peking University People's Hospital
Cuizhi Geng
The Fourth Hospital of Hebei Medical University
Yi Li
Shunyi District Health Care Hospital for women and Children of Beijing
XiangJun Ma
Haidian Maternal and Child Health Hospital
Bin Wang
Peking University First Hospital
YingJian He
Peking University Cancer Hospital: Beijing Cancer Hospital
Chen Yao (yaochen@hsc.pku.edu.cn)
Peking University https://orcid.org/0000-0002-4916-4204
Tao Ouyang
Peking University Cancer Hospital: Beijing Cancer Hospital

Research article

Keywords: Breast cancer, Benign, Machine learning model, Malignant, Diagnosis

Posted Date: November 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-101184/v1

Abstract

Background: The accuracy of breast cancer (BC) screening based on conventional ultrasound imaging examination largely depends on the experience of clinicians. Further, the effectiveness of BC screening and diagnosis in primary hospitals need to be improved. This study aimed to establish and evaluate the usefulness of a simple, practical, and easy-to-promote machine learning model based on ultrasound imaging features for diagnosing BC.

Methods: Logistic regression, random forest, extra trees, support vector, multilayer perceptron, and XG boost models were developed. The modeling data set was divided into a training set and test set in a 75%:25% ratio, and these were used to establish the models and test their performance, respectively. The validation data set of primary hospitals was used for external validation of the model. The area under the receiver operating characteristic curve (AUC) was used as the main evaluation index, and pathological biopsy was used as the gold standard for evaluating each model. Diagnostic capability was also compared with those of clinicians.

Results: Among the six models, the logistic model showed superior capability, with an AUC of 0.771 and 0.906 in the test and validation sets, respectively, and Brier scores of 0.18 and 0.165. The AUC of the logistic model in tertiary class A hospitals and primary hospitals was 0.875 and 0.921, respectively. The AUCs of the clinician diagnosis and the logistic model were 0.913 and 0.906. Their AUCs in the tertiary class A hospitals were 0.890 and 0.875, respectively, and were 0.924 and 0.921 in primary hospitals, respectively.

Conclusions: The logistic regression model has better overall performance in primary hospitals, and the logistic regression model can be further extended to the basic level. A more balanced clinical prediction model can be further established on the premise of improving accuracy to assist clinicians in decision making and improve diagnosis.

Trial Registration: http://www.clinicaltrials.gov. ClinicalTrials.gov ID: NCT03080623.

Background

Breast cancer (BC) is the most common malignancy among women worldwide [1–3]. However, the exact cause of BC is yet to be fully understood. BC screening for early diagnosis is crucial for improving treatment efficacy and survival [4]. Many previous studies have confirmed that early detection significantly increases the probability of survival by preventing metastasis [5]. In Beijing, the primary method of BC screening is breast ultrasound imaging examination. However, given that the accuracy of conventional ultrasound imaging is highly dependent on the clinicians' expertise and experience, the results of BC screening and diagnosis in primary hospitals are suboptimal [6]. Therefore, a model for diagnosing breast lesions based on the characteristics of large samples of ultrasound images may be helpful for lowering subjectivity and improving the accuracy of screening. Currently, only a logistic regression model is used for establishing diagnostic models for manual interpretation of ultrasound findings. The evaluation model is often verified by internal verification, which may overestimate the performance of the models [7–9].

Subjects And Methods

Aims

This study aimed to establish a simple, practical, and easy-to-promote clinical model for BC diagnosis and evaluate its usefulness in primary hospitals. Towards this goal, we screened out meaningful predictors based on the data collected by tertiary class A hospitals and established diagnostic models. Population data, including from primary hospitals, were used as an external verification data set to validate the effectiveness of the model and explore its applicability and clinical potential.

Data sets

The modeling data set was a cumulative collection of data from 1345 patients admitted to Beijing Cancer Hospital between November 2010 and May 2016. Data on ultrasound findings and histopathological diagnosis were collected. Ultrasound was performed using a SomoV automatic breast ultrasound system from the U-system. The probe frequency was 10 MHz, and the probe size was 15 cm×17 cm×5 cm. The maximum diameter of the ultrasound image of the lesion was less than 2 cm. Two-dimensional images were collected, and the coronal image was reconstructed. After re-evaluation by professional cliniciansfrom Beijing People's Hospital, the cases with consistent findings were selected as the final modeling data set. In total, data from 1125 patients were included; of them, 732 patients had malignant tumors.

The external verification data set was from five centers, namely, Beijing Cancer Hospital, Beijing People's Hospital, Fourth Hospital of Hebei Medical University, Beijing Shunyi District Maternity and Child Health Hospital, and Beijing Haidian District Maternity and Child Health Hospital. The data were cumulatively collected from August 2017 to December 2019 and comprised pathological results of 1981 biopsy (n=1094) or follow-up (n=890) cases. After data cleaning, 1965 cases were included in the verification data set.

The dependent variable of the machine model was the diagnosis result (benign or malignant) of biopsy cases with pathological biopsy classification or follow-up cases with disease classification.

The independent variable was the expert group and modeling working group classification from Peking University Cancer Hospital and Peking University People's Hospital. This working group extracted and clarified the definitions of ultrasound imaging terminology based on the interpretation of ultrasound images in a blinded manner. We have previously published relevant literature [10] using the full model strategy, logistic model strategy, and random forest model strategy to screen independent variables and establish models.

The external validation data set comprised only part of the screening independent variables that need to be validated based on the previous models. The identifiable information of the boundary was classified into 4 features when the boundary was not identifiable. The specific variable assignments are shown in Table 2.

Methods

The data set was divided into a modeling data set and an external verification dataset. We selected 75% of the samples from the modeling data set as the training set. The variable selection, one-hot encoding, and basic model were assembled into a pipeline, which was placed enter the grid search, using the 10-fold cross validation technique. In this technique, the data set was divided into 10 folds, and each fold was used for internal verification. The remaining 90% was used for the training of the development model. The hyperparameter adjustment was used for establishing the model. Otherwise, we validated the models with the remaining 25% of the samples and external validation data sets. Cross-validation and hyperparameter adjustments for internal validation are considered robust methods of model evaluation before external validation on a separate data set. This could maximize the potential performance of machine learning models [11-15].

We validated each model through an external verification data set. The discriminative capability of each model was validated using the area under the receiver operating characteristic (ROC) curve. Meanwhile, the Brier score was calculated to quantify the calibration degree of the model, and a calibration degree scatter diagram was created thereafter. We then evaluated the consistency of the actual observations and models according to the comparison between the scattered point distribution and the reference line.

The verification data were stratified according to primary hospitals (the Fourth Hospital of Hebei Medical University, Beijing Shunyi District Maternity and Child Health Hospital, and Beijing Haidian District Maternity and Child Health Hospital) and Beijing tertiary class A hospitals (Beijing Tumor Hospital and Beijing People's Hospital) to compare between each model and the results determined by clinicians.

Statistical analysis

Raw data were cleaned using SAS v.9.4(SAS Institute, Cary, NC) and a single factor analysis was performed. The categorical independent and dependent variables were evaluated using chi-square test. The verification process was mainly based on the third-party "Sklearn" library of Python (version 0.22.2.post1). The area under the AUCwas calculated to assess the model discrimination. The AUC value ranges from 0.5 to 1, and the closer the AUC is to 1, the better the discriminative capability of the model. An AUC of 0.5 indicates that the model is not predictive and has no practical application. We evaluated the model calibration using the Brier score and calibration curve. The Brier score is calculated using the formula $(Y\text{-}p)^2$, where Y is the actually observed outcome variable (0 or 1), and p is the predicted probability based on the prediction model. The Brier score ranges from 0 to 0.25, and the smaller the score, the better the calibration of the model. A Brier score of 0.25 indicates that the model has no predictive capability [16-17].

Results

Basic information

The modeling data set included data from 732 cases of malignant tumors (65.07%) and 393 cases of benign tumors (34.93%). Meanwhile, the validation data set included data from 498 cases of malignant tumors (25.34%) and 1467 cases of benign tumors (74.66%). With respect to clinician findings in the validation data set, 1354 follow-up cases (68.91%) and 611 biopsy cases (31.09%) were determined to be

malignant, respectively. Pathological examination of the biopsy cases revealed 498 malignant tumors (45.69%) and 592 benign tumors (54.31%). All follow-up cases were benign tumors (100%) on pathological examination. Comparison of the predictive variables between the modeling data set and validation data set showed a significant difference in the distribution of these predictors (P<0.001, Table 3).

Comparison between benign and malignant tumors

Univariate analysis of the independent variables in the validation data set identified seven predictors, namely, direction, margin blur, margin angulation, margin microlobulation, margin burr, posterior echoes, and surrounding tissue edema. Further, their distribution was significant different between the benign and malignant groups (P<0.001).

Discriminative capability of the machine learning models

The degree of discrimination was used to evaluate the discriminative and ranking capabilities of the model, which indicate the model's capability to distinguish between individuals with end-point events and individuals without end-point events [18]. In the internal verification, there were no significant differences in the results of several models after hyperparameter adjustment. The multilayer perceptron model performed best, with an AUC (95% CI) of 0.782 (0.724-0.835). In the external verification, the logistic regression model performed best after hyperparameter adjustment, with an AUC (95% CI) of 0.906 (0.892-0.921). The performance of the model in the verification set was generally better than that in the test set. The indicators of each model are shown in Table 5, and the ROC curves are shown in Figure 1.

Calibration of the machine learning models

Compared with discrimination, calibration pays more attention to the accuracy of the absolute risk prediction value of the model, that is, the consistency between the probability of the outcome predicted by the model and the probability of the actual outcome [18]. In the internal verification, the Brier scores of the logistic regression, random forest, extra trees, support vector, multilayer perceptron, and XG boost were 0.181, 0.189, 0.196, 0.199, 0.177, and 0.179, respectively. In the external verification, logistic regression, random forest, extra trees, support vector, multilayer perceptron, and XG boost were 0.165, 0.163, 0.170, 0.178, 0.146, and 0.161, respectively. The calibration curves are shown in Figure 2.

Comparison of outcomes between clinician and models

We compared the predicted outcome of the models with the results determined by clinicians according to the center stratification (Table 6). Overall, clinician diagnosis showed a higher accuracy than model diagnosis. The model had an accuracy of 0.906; sensitivity, 0.928; specificity, 0.898; and AUC, 0.913. The accuracy of clinician diagnosis in primary hospitals was 0.929; the AUC was 0.924; and the sensitivity and specificity were 0.918 and 0.930, respectively. The accuracy of clinician diagnosis in the tertiary class A hospitals was 0.880; the AUC was 0.849; and the sensitivity and specificity were 0.932 and 0.898. When comparing clinician diagnosis between primary and tertiary class A hospitals, the sensitivity was higher in the tertiary class A hospitals, while the accuracy, specificity, and AUC were lower than those in the primary

hospitals. Further, we found that each model had a better predictive performance among patients in primary hospitals than those in tertiary class A hospitals (AUC: 0.921 vs. 0.875, Table 7).

Discussion

Based on our previous study that initially identified 27 independent variables [10], we selected 7 independent variables, namely, direction, margins blur, margins angulation, margins microlobulation, margins burr, posterior echoes, and surrounding tissue edema, in this study to develop six machine learning models for BC diagnosis. The logistic model showed superior performance, with an ROC of 0.771 and 0.906 in the test set and the validation set and Brier scores of 0.18 and 0.165, respectively. As such, we recommend using a logistic regression model fitted with ultrasound imaging features for BC diagnosis, particularly in primary hospitals.

Logistic regression can identify important predictors of BC using odds ratios and generate confidence intervals that provide additional information for decision-making [19]. In our logistic regression model, tumor margins burr and the direction of tumor growth had a relatively large impact on the judgment of benign and malignant tumors. The odds ratio (OR) were 3.267 (2.013–5.303) and 4.281 (3.098–5.917), respectively. This is consistent with the findings reported by Chhatwal et al. [20] that the most important predictors associated with BC as identified by this model were spiculated mass margins. In the current study, the OR value represents the ratio of the risk of malignant BC based on the existence and absence of a certain ultrasound feature. The greater the OR value (OR > 1), the greater the risk of malignancy in the presence of the feature. Direction of tumor growth, non-identifiable and burr at the margins, and edema of the surrounding tissue showed the highest OR values, indicating that non-parallel growth, non-identifiable margins burr, and edema of the surrounding tissue are the most important factors for predicting malignant BC. This is consistent with the findings of previous studies. Nianan [21] reported that non-parallel growth and irregular morphology are the most important predictors of BC in the new version of the BI-RADS. Some studies have also shown that axillary lymphadenopathy is indicative of the probability of metastasis in BC [22–23].

The average AUC of models in the test set was 0.741 ± 0.052, and the average AUC in the validation set was 0.880 ± 0.025. The overall performance of the model in the validation set was better than that in the test set. Compared with internal verification, external verification is more concerned with model transportability and generalizability [24–26]. Thus, we believe that the predictive model can be applied generally across population samples and has good promotion significance.

When compared with clinician diagnosis, the logistic regression model showed lower accuracy (0.906 vs. 0.772) and AUC (0.913 vs. 0.906). When model performance was evaluated by type of hospital (tertiary class A hospitals and primary hospitals), the model performed better in primary hospitals than in tertiary class A hospitals. This may be due to the different distribution of benign and malignant tumors in both groups. The proportion of benign tumor patients was significantly higher in primary hospitals (n = 892, 85.93%) than that in tertiary class A hospitals (n = 575, 62.02%). For complex malignant tumors, predictions based on models alone is more likely to be biased. In primary hospitals, the accuracy of clinician diagnosis

was higher than that of the logistic model (0.929 vs. 0.806), and the AUC of clinician diagnosis was also slightly higher (0.913 vs. 0.906). Similarly, the accuracy of clinician diagnosis in tertiary class A hospitals was higher than that of the logistic model (0.880 vs. 0.734). The AUC of clinician diagnosis was also slightly higher than that of the logistic model (0.890 vs. 0.875). The high sensitivity of clinician diagnosis in tertiary class A hospitals indicates that clinicians have a greater probability of accurately diagnosing malignant tumors, and the possibility of missed diagnosis is lower. Meanwhile, the high specificity of clinician diagnosis in primary hospitals indicates that clinicians in these hospitals can accurately diagnosis benign tumors, and the possibility of misdiagnosis is lower. Although there was no significant difference in AUC between the models and clinician diagnosis, the accuracy was markedly different. This may be caused by the imbalance in the distribution of samples between the malignant group and the benign group. Subsequent studies should validate the usefulness of the model by using equally distributed samples, particularly in primary hospital population alone. This will ultimately help establish the use of the model in primary hospitals.

Our models enable the prediction of BC and can thus be used by clinicians to make appropriate patient management decisions. As shown in Fig. 3, the predictive capability of the models ranged from 0.2 to 0.4. We analyzed the model prediction probabilities according to 1%, 2%, 5%, 10%, 50%, 90%, 95%, 98%, and 99% and applied the logistic model in the clinic for preliminary evaluation of BC. If the predicted probability was lower than 1% of the population (corresponding to a predicted probability of 0.2158926), it is highly likely that patients do not have to undergo pathological biopsy. Malignancy can be largely ruled out, and the patient can undergo regular follow-up. When the predicted probability is higher than 90% of the population (corresponding to a predicted probability of 0.8769365), it is highly indicative of malignant lesions, and clinicians are required to intervene. Patients should immediately undergo a pathological biopsy to confirm malignancy. For patients whose predicted probabilities are in between these values, a short-term follow-up (within 1 year, preferably 3 to 6 months) can be recommended [27]. The clinicians can further use the models to assist in decision-making according to the follow-up outcomes. However, the cut-off value of the predictive probability needs to be verified and calculated in studies with a larger sample size to improve the accuracy.

This study has some limitations. First, this study was mainly an external verification of the previous model. The independent variable in the model population is different from the verification population, which may cause a selection bias. Second, this study did not modify and improve the model because of the imbalance in the distribution of the predictor variables and classification, and thus the model has low accuracy. Future studies should take measures to account for accuracy in the modeling process. Third, this study did not collect demographic information and baseline data of the patients, and it was difficult to balance the patient baseline in the pre-modeling stage. This may have affected the performance of the model and introduced confounding factors. Further studies are needed to improve model accuracy and to establish a more balanced clinical prediction model that can be used not only during diagnosis, but also at follow-up.

In conclusion, of the six machine learning models, the logistic regression model showed the highest predictive capability and generalizability, indicating its potential for application in primary hospitals. The model showed similar predictive performance to clinicians. Further, it had better predictive capability in

primary hospitals than in tertiary class A hospitals model. Collectively, these findings indicate that the model can help clinicians in distinguishing between benign and malignant breast tumors.

List Of Abbreviations

AUC, area under the receiver operating characteristic curve

BC, breast cancer

OR, odds ratio

Declarations

Ethics approval and consent to participate

Ethical approval of the study protocol was granted by the Ethics Committe of Beijing Cancer Hospital (Approval Number: 2016KT14) in Beijing, China. All patients provided written informed consent to participate in this study

Consent for publication: Not applicable.

Availability of data and materials:

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by Beijing Municipal Science and Technology Project (NO: D161100000816006).

Authors' contributions

Yao Tan contributed to development of the models, data cleaning and manuscript writing, mainly contributed to the article methodology. Ling Huo provided clinical ideas and discussion content, assisted article writing and revision. Bin Wang assisted model construction and chart generation. YingJian He was responsible for coordinating various centers for data collection and aggregation. Tao Ouyang was the Principal Investigator in charge of the project, designed, supervised the project, and Chen Yao supervised the study and provided analysis method ideas. Tao Ouyang and Chen Yao gave the final approval for this version of the manuscript to be published and agreed to be accountable for all aspects of this work. All authors reviewed the final version of the manuscript and approved it for publication.

Acknowledgements

We are grateful to all patients and would also like to thank all of the staff involved in the study at the participating study sites (5 sites).

1)Department of Breast Center, Peking University People's Hospital, Beijing, People's Republic of China; Shu Wang, email: shuwang@pkuph.edu.cn.

2) Haidian maternal and child health hospital; XiangJun Ma, email: xiangjunma1962@126.com.

3) Shunyi District Health Care Hospital for Women and Children of Beijing; Yi LI, email: Liyiborui@126.com.

4) The Fourth Hospital of Hebei Medical University; CuiZhi Geng, email: gengcuizhi@hotmail.com.

Authors' information (optional)

[1] Peking University First Hospital, Beijing, China.

[2] Key laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Breast Center, Peking University Cancer Hospital & Institute

[3] Peking University Clinical Research Institute, Peking University Health Science Center, Beijing, China;

[4] Department of Breast Center, Peking University People's Hospital, Beijing, People's Republic of China.

[5] The Fourth Hospital of Hebei Medical University.

[6] Shunyi District Health Care Hospital for Women and Children of Beijing.

[7] Haidian maternal and child health hospital.

References

1. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66(2):115–32.
2. Stalin M, Kalaimagal R. Breast cancer diagnosis from low-intensity asymmetry thermogram breast images using fast support vector machine, i-manager's Journal on Image Processing. 2016:17–26.
3. Kirubakaran R, Jia TC, Aris NM. Awareness of breast cancer among surgical patients in a tertiary hospital in Malaysia. Asian Pac J Cancer Prev. 2017;18:115–20.
4. Lee C. Delay in breast cancer: Implications for stage at diagnosis and survival. Front Public Health. 2014;2:87.
5. Richards MA, Westcombe AM, Love SB, Littlejohns P, Ramirez AJ. Influence of delay on survival in patients with breast cancer: a systematic review. Lancet. 1999;353:1119–26.

6. Lili H, Qingqing Q, Chao W, et al. Discussion on the main problems and experience of two cancer screening in Beijing. Practical Preventive Medicine. 2011;018:566–8.
7. Zhang Y, Zhao Z, Cai J, et al. Study on logistic regression analysis and risk prediction model establishment with ultrasound image features for differential diagnosis of breast neoplasms. Journal of Chinese Oncology. 2016;22:214–7.
8. Yun Wu, Hong L. Correlation analysis of breast ultrasound imaging features and malignant breast masses. Medical Journal of National Defending Forces in Southwest China. 2020;30:47–9.
9. Zhao H, Peng Y, Luo H, He Y, et al. A logistic regression model based on breast imaging report and data system lexicon to predict the risk of malignancy. West China Medical Journal. 2015;30:2249–53.
10. Zilong Z, Yingjian H, Tao O, et al. Application value of random forest and support vector machine in diagnosing breast lesions by using ultrasonic image features. Chinese Journal of Health Statistics. 2018;35:684–8.
11. Steyerberg EW. Validation in prediction research: the waste by data splitting. J Clin Epidemiol. 2018;103:131–3.
12. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J. 2017;38:500–7.
13. Samad MD, Ulloa A, Wehner GJ, Jing L, Hartzel D, Good CW, et al. Predicting survival from large echocardiography and electronic health record datasets: optimization with machine learning. JACC Cardiovasc Imaging. 2019;12:681–9.
14. Kennedy EH, Wiitala WL, Hayward RA, Sussman J. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med Care. 2013;51:251–8.
15. Rotondano G, Cipolletta L, Grossi E, Koch M, Intraligi M, Buscema M, et al. Artificial neural networks accurately predict mortality in patients with nonvariceal upper GI bleeding. Gastrointest Endosc. 2011;73:218–26.
16. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. Berlin: Springer Science &Business Media; 2008.
17. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38.
18. Wang J, Zhang Z, Zhou Z, Gu H. Clinical prediction models: Model validation. Chin J Evid Based Cardiovasc Med. 2019;11:141–4.
19. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models:seven steps for development and an ABCD for validation [J].Eur Heart J,2014,35(29):1925–31.
20. Chhatwal J, Alagoz O, Lindstrom MJ, Kahn CE, Shaffer KA, Burnside ES. A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. Am J Roentgenol. 2009;192:1117–27.
21. Nianan H. 2013 edition of breast ultrasound report and data system classification interpretation and new progress in clinical application. Anhui Medical Journal. 2015;36(11):1424–7.

22. Wang ZL, Li JL, Li M, Huang Y, Wan WB, Tang J. Study of quantitative elastography with supersonic shear imaging in the diagnosis of breast tumours. Radiol Med. 2013;118:583–90.
23. Meng W, Ru-hai Z. Values of breast ultrasonography using BI-RADS classification in differentiating benign and malignant breast masses. Journal of China Clinic Medical Imaging. 2014;(6):390–392.
24. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691–8.
25. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98:683–90.
26. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal external, and external validation. J Clin Epidemiol. 2016;69:245–7.
27. Breast Imaging Reporting And Data System (BI-RADS). Vol. 4. Reston VA: American College of Radiology; 2004.

Tables

Table 1
AUC of the independent variable screening strategy models

Strategies	Logistic regression (95% CI)	Random forest (95% CI)
Full models	0.7812(0.7325–0.8298)	0. 7878(0.7392–0.8365)
Logistic	0.7727 (0.7227–0.8227)	0. 7757 (0.7258–0.8255)
Random forest	0.7880 (0.7395–0.8364)	0. 7868 (0.7377–0.8359)

Table 2
Variable assignment

Variable	Name	Value
Breast left/right	zyc	0-left,1-right
Direction	FX	0- parallel,1-unparallel
Margins blur	bqxcd1	0-identifiable, 1-non-identifiable but no blur, 2-non-identifiable and blurred
Margins angulation	bqxcd2	0-identifiable, 1-non-identifiable but no angulation, 2-non-identifiable and angled
Margins microlobulation	bqxcd3	0-identifiable, 1-non-identifiable but no microlobulation, 2-non-identifiable and microlobulated
Margins burr	bqxcd4	0-identifiable, 1-non-identifiable but no burr, 2-non-identifiable and burr
Posterior echoes	hfhs	0-no change, 1-enhanced, 2- attenuated(include mixed)
Surrounding tissue edema	shuiz	0-no, 1-yes
Benign vs. malignant	end	0- benign, 1-malignant
Clinicians	biras	0-benign tendency (follow-up), 1- malignant tendency (biopsy)
Biopsy results	path	0- benign,1-malignant
Follow-up results	path3	0- benign,1-malignant

Table 3
Comparison between the modeling data set and the validation data set

Variable		Modeling data set (n = 1125)	Validation data set (n = 1965)	χ^2	*P*
Zyc	left n(%)	0 (0.00%)	942 (47.94%)	-	-
	right n(%)	0(0.00%)	1023(52.06%)		
FX	Parallel	826 (73.42%)	1566 (79.69%)	16.096	0.000
	Unparallel	299 (26.58%)	399 (20.31%)		
Bqxcd1	Identifiable	160 (14.22%)	1074 (54.66%)	609.309	0.000
	Non-identifiable but no blur	80(7.11%)	240(12.21%)		
	Non-identifiable and blurred	885 (78.67%)	651 (33.13%)		
Bqxcd2	Identifiable	160 (14.22%)	1073 (54.61%)	504.371	0.000
	Non-identifiable but no angulation	525 (46.67%)	401 (20.41%)		
	Non-identifiable and angled	440 (39.11%)	491 (24.99%)		
Bqxcd3	Identifiable	160 (14.22%)	1073 (54.61%)	629.396	0.000
	Non-identifiable but no microlobulation	363 (32.27%)	574 (29.21%)		
	Non-identifiable and microlobulated	602 (53.51%)	318 (16.18%)		
Bqxcd4	Identifiable	160 (14.22%)	1074 (54.66%)	497.430	0.000
	Non-identifiable but no burr	720 (64.00%)	717 (36.49%)		
	Non-identifiable and burr	245 (21.78%)	174 (8.85%)		
hfhs	No change	687 (61.07%)	1549 (78.83%)	114.225	0.000
	Enhanced	198 (17.60%)	204 (10.38%)		
	Attenuated (including mixed)	240 (21.33%)	212 (10.79%)		
shuiz	No	1079 (95.91%)	1823 (92.77%)	12.326	0.000
	Yes	46 (4.09%)	142 (7.23%)		
End	Benign	393 (34.93%)	1467 (74.66%)	471.132	0.000

The values are presented in n (%).

Variable		Modeling data set (n = 1125)	Validation data set (n = 1965)	χ^2	*P*
	Malignant	732 (65.07%)	498 (25.34%)		
The values are presented in n (%).					

Table 4
Comparison between the benign and malignant groups in the validation set

Variable		Benign (n = 1467)	Malignant (n = 498)	χ^2	*P*-value
Zyc	Left	1352(92.16%)	214(42.97%)	555.895	0.000
	Right	115(7.84%)	284(57.03%)		
FX	Parallel	1040(70.89%)	34(6.83%)	656.956	0.000
	Unparallel	152 (10.36%)	88 (17.67%)		
Bqxcd1	Identifiable	275 (18.75%)	376 (75.50%)		
	Non-identifiable but no blur	1040 (70.89%)	33 (6.63%)	657.869	0.000
	Non-identifiable and blurred	232 (15.81%)	169 (33.94%)		
Bqxcd2	Identifiable	195 (13.29%)	296 (59.44%)		
	Non-identifiable but no angulation	1040 (70.89%)	33 (6.63%)	679.549	0.000
	Non-identifiable and angled	323 (22.02%)	251 (50.40%)		
Bqxcd3	Identifiable	104 (7.09%)	214 (42.97%)		
	Non-identifiable but no microlobulation	1040 (70.89%)	34 (6.83%)	808.091	0.000
	Non-identifiable and microlobulated	415 (28.29%)	302 (60.64%)		
Bqxcd4	Identifiable	12 (0.82%)	162 (32.53%)		
	Non-identifiable but no burr	1271 (86.64%)	278 (55.82%)	231.661	0.000
	Non-identifiable and burr	116 (7.91%)	88 (17.67%)		
hfhs	No change	80 (5.45%)	132 (26.51%)		
	Enhanced	1440 (98.16%)	383 (76.91%)	250.462	0.000
	Attenuated (include mixed)	27 (1.84%)	115 (23.09%)		
The values are presented in n (%)					

Table 5
Performance evaluation of the different models

Model	Accuracy	Precision class 1	Recall class 1	AUC of ROC	AUC of PRC	F1 score
Test set (calibration model)						
Logistic Regression	0.720	0.734	0.891	0.771	0.846	0.805
Random forest	0.727	0.755	0.858	0.747	0.812	0.803
Extra trees	0.723	0.754	0.852	0.746	0.820	0.800
Support vector	0.709	0.717	0.913	0.638	0.736	0.803
Multilayer Perceptron	0.738	0.756	0.880	0.775	0.838	0.813
XG Boost	0.713	0.730	0.885	0.769	0.839	0.800
Validation set (calibration model)						
Logistic Regression	0.772	0.528	0.936	0.906	0.794	0.675
Random forest	0.814	0.598	0.813	0.865	0.735	0.689
Extra trees	0.813	0.597	0.807	0.855	0.709	0.687
Support vector	0.768	0.524	0.936	0.852	0.632	0.671
Multilayer Perceptron	0.818	0.596	0.869	0.901	0.792	0.708
XG Boost	0.781	0.542	0.876	0.898	0.776	0.669

Table 6
Comparison between clinician diagnosis and gold standard diagnosis

Clinicians		Gold standard		Total
		Benign	Malignant	
All validation set	Benign	1318	36	1354
	Malignant	149	462	611
	Total	1467	498	1965
Primary hospitals	Benign	830	12	842
	Malignant	62	134	196
	Total	892	146	1038
Tertiary class A hospitals	Benign	488	24	512
	Malignant	87	328	415
	Total	575	352	927

Table 7
Comparison between clinician and model diagnosis

Model	Accuracy	Precision class 1	Recall class 1	AUC of ROC	AUC of PRC	F1 score	Threshold	FPR	TPR
Full validation set									
Clinicians	0.906	0.756	0.927	0.913	0.851	0.833	-	-	-
Logistic regression	0.772	0.528	0.936	0.906	0.794	0.675	0.571	0.181	0.829
Random forest	0.814	0.598	0.813	0.865	0.735	0.689	0.491	0.185	0.815
Extra trees	0.813	0.597	0.807	0.855	0.709	0.687	0.505	0.185	0.807
Support vector	0.768	0.524	0.936	0.852	0.632	0.671	0.71	0.206	0.793
Multilayer perceptron	0.818	0.596	0.869	0.901	0.792	0.708	0.573	0.187	0.827
XG boost	0.781	0.542	0.876	0.898	0.776	0.669	0.557	0.183	0.817
Primary hospitals									
Clinicians	0.929	0.683	0.918	0.924	0.807	0.784	-	-	-
Logistic regression	0.806	0.416	0.932	0.921	0.768	0.575	0.566	0.146	0.836
Random forest	0.884	0.560	0.829	0.910	0.725	0.669	0.48	0.115	0.877
Extra trees	0.885	0.562	0.836	0.904	0.703	0.672	0.505	0.107	0.836
Support vector	0.804	0.413	0.932	0.887	0.560	0.573	0.702	0.204	0.836
Multilayer perceptron	0.882	0.549	0.884	0.923	0.779	0.677	0.52	0.119	0.884
XG boost	0.829	0.447	0.904	0.926	0.767	0.599	0.537	0.118	0.884
Tertiary class A hospitals									
Clinicians	0.880	0.790	0.932	0.890	0.874	0.855	-	-	-
Logistic regression	0.734	0.595	0.938	0.875	0.818	0.728	0.627	0.209	0.787
Random forest	0.736	0.616	0.807	0.814	0.750	0.699	0.565	0.285	0.764
Extra trees	0.732	0.614	0.795	0.802	0.722	0.693	0.567	0.285	0.753

Model	Accuracy	Precision class 1	Recall class 1	AUC of ROC	AUC of PRC	F1 score	Threshold	FPR	TPR
Support vector	0.727	0.588	0.938	0.787	0.665	0.723	0.733	0.297	0.705
Multilayer perceptron	0.746	0.619	0.864	0.861	0.803	0.721	0.601	0.207	0.767
XG boost	0.726	0.596	0.864	0.854	0.787	0.705	0.587	0.205	0.756

Table 8
Performance of the logistic regression model

	B	SE	OR	95% *CI*	*P*	β
fx	1.454	0.165	4.281	3.098–5.917	< 0.001	0.322239
bqxcd1	0.235	0.143	1.265	0.956–1.674	0.100	0.118155
bqxcd2	0.334	0.142	1.396	1.058–1.844	0.019	0.155041
bqxcd3	0.716	0.154	2.047	1.513–2.768	< 0.001	0.295653
bqxcd4	1.184	0.247	3.267	2.013–5.303	< 0.001	0.425586
hfhs	0.340	0.101	1.405	1.152–1.714	0.001	0.123337
shuiz	1.193	0.269	3.298	1.947–5.586	< 0.001	0.170345

Table 9
Predicted probability of different proportions of people by model

	Logistic regression	Random forest	Extra trees	SVC	MLP classifier	XGB classifier
1%	0.2158926	0.0870467	0	0.2690289	0.1223317	0.1271728
2%	0.2481656	0.2063348	0.1830000	0.2690872	0.1924786	0.2399745
5%	0.2953400	0.2432472	0.2500000	0.2691355	0.2032477	0.2851146
10%	0.2953400	0.2826738	0.2857143	0.2691355	0.2580033	0.2999176
50%	0.2953400	0.2826738	0.2857143	0.2691355	0.2580033	0.2999176
90%	0.8769365	0.8999733	0.9291429	0.7422661	0.8754747	0.8494976
95%	0.9327307	0.9831579	1	0.7428197	0.9669854	0.9255747
98%	0.9648594	1	1	0.7554798	0.9834885	0.9730366
99%	0.9675776	1	1	0.7882681	0.9877369	0.9751260

Figures

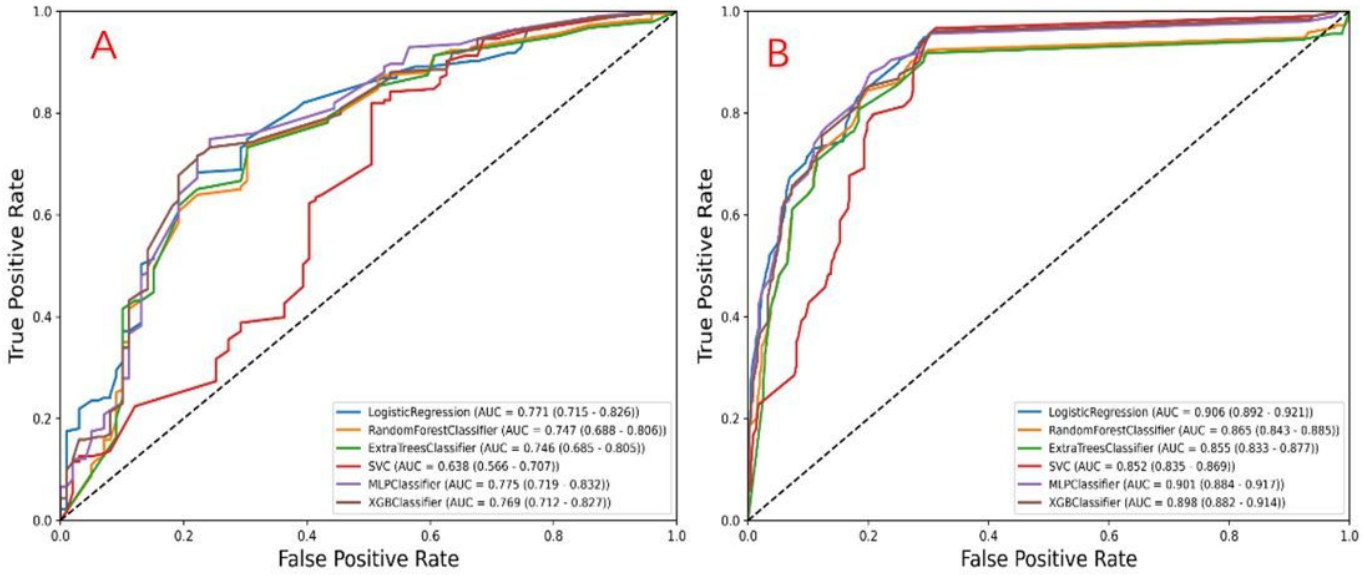

Figure 1

ROC plots of the calibrated model in the test set (A) and validation set (B).

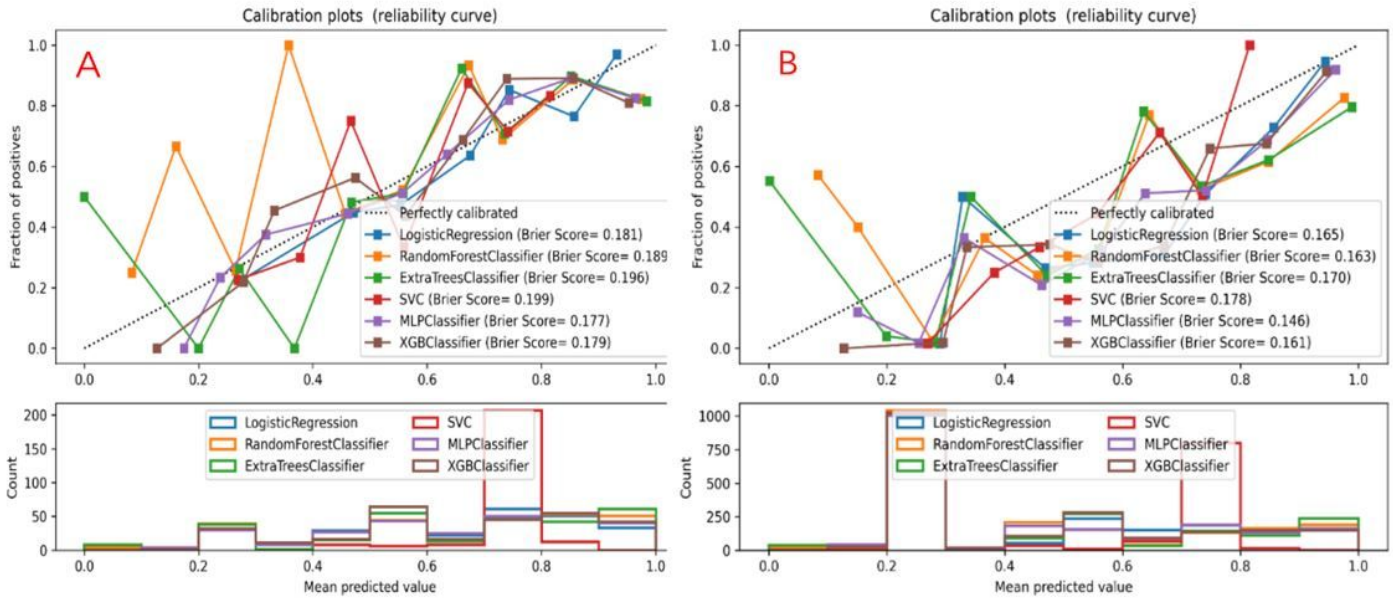

Figure 2

Calibration plots of the calibrated model in the test set (A) and validation set (B).

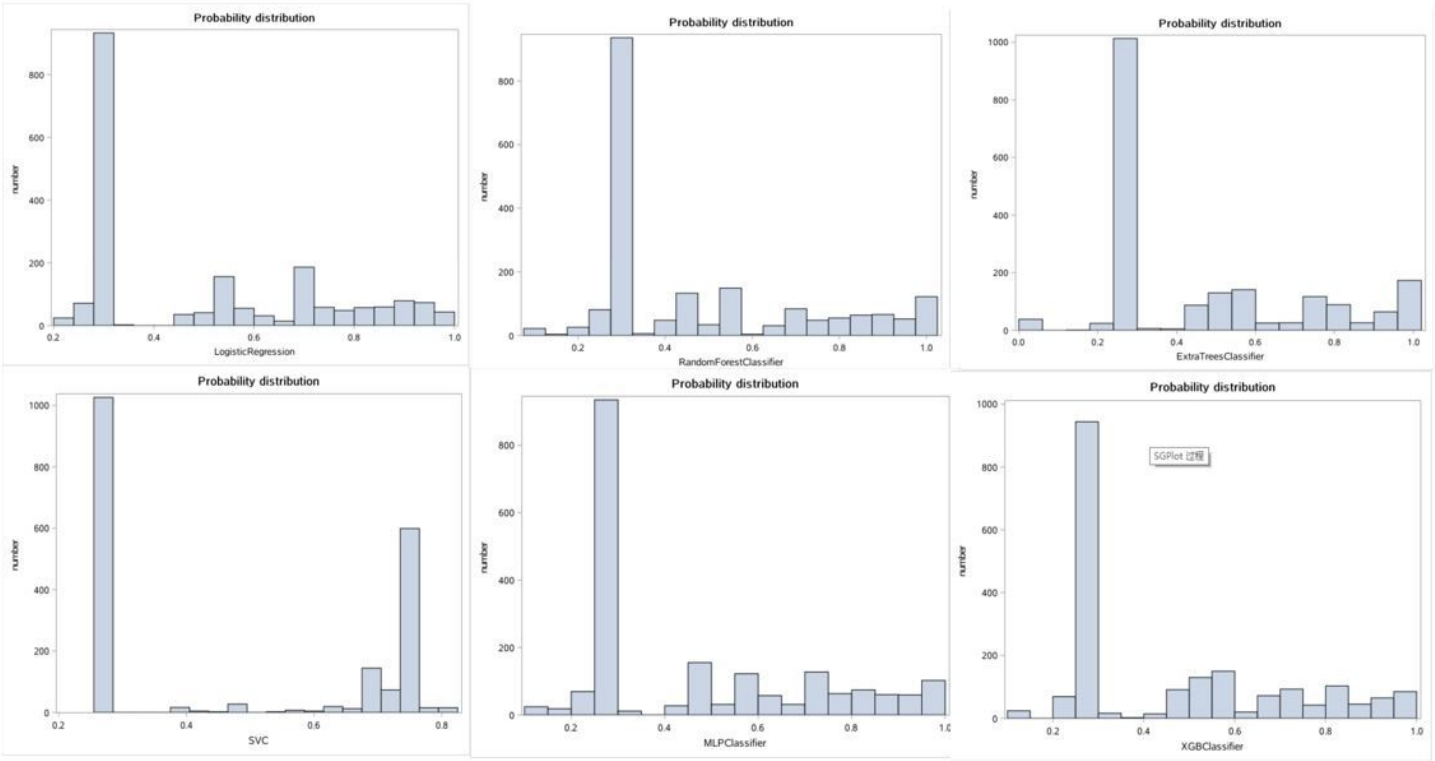

Figure 3

Probability distribution by model.