Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal Ultrasound Combined with Machine Learning for Diagnosing Breast Lesions: Incremental Value Assessment of Ultrasound Super-Resolution Imaging

Bin Shi
The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China

Yang Yang
The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China

Jing Zhang
The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China

Yun Yuan
The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China

Chao Wei
The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China

Kun Tao
The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China

Lei Ye
yeleius@ustc.edu.cn

The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China

---

Research Article

Keywords: Breast Lesions, Multimodality, Ultrasound Super-Resolution Imaging, Machine Learning

Posted Date: August 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10399234/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background. To integrate conventional ultrasound, contrast-enhanced ultrasound (CEUS), and ultrasound super-resolution imaging (US SRI) into an interpretable machine learning model to distinguish benign from malignant breast lesions.

Methods. We retrospectively enrolled 210 cases (106 benign, 104 malignant). Following feature selection, six models were employed: Logistic Regression, Decision Tree, Random Forest (RF), Extreme Gradient Boosting, Light Gradient Boosting Machine, and Support Vector Machine. Model stability and generalization were evaluated using 10-fold cross-validation (CV), an independent test set, and prospective validation (53 cases: 18 benign, 35 malignant). Performance was evaluated with the area under the receiver operating characteristic curve (AUC), calibration curves, and decision curve analysis (DCA), with SHapley Additive exPlanations (SHAP) for interpretability.

Results. Lasso selected eight features: age, minimum diameter, ratio of diameter, anatomical location, breast imaging reporting and data system (BI-RADS) on CEUS, directional variance (Dir.Var), 0.5mm peritumoral perfusion index (0.5pPI), and 1.0mm peritumoral vascular density (1.0pVD). In the test set, the RF model performed best (AUC = 0.9909). SHAP identified BI-RADS on CEUS, age, and minimum diameter as core clinical features, while the three US SRI parameters provided minor but non-negligible incremental information. Calibration curves and DCA demonstrated good agreement and excellent clinical net benefit. In the prospective cohort, the RF model achieved an AUC of 0.9492, significantly higher than that of junior ultrasound physicians ($P < 0.05$).

Conclusion. The multimodal machine learning model that integrated macroscopic and microscopic features exhibited excellent generalization and clinical utility for breast lesion differentiation.

Background

Breast cancer is the most common and deadliest malignancy among women worldwide [1]. Its pathogenesis is multifactorial, involving genetic, hormonal, reproductive, lifestyle, and environmental factors [2]. Early differentiation between benign and malignant breast lesions is crucial, as it guides the choice between aggressive surgical intervention and conservative follow-up [3].

Conventional grayscale ultrasound and color Doppler flow imaging (CDFI) can assess lesion morphology, location, echo characteristics, and blood flow, but their features overlap considerably between benign and malignant breast lesions [4–5]. CEUS improves diagnostic sensitivity and specificity by visualizing tumor perfusion, yet it cannot resolve microvascular details due to the diffraction limit (microvessels are < 100 μm, while CEUS resolution is limited to >half a wavelength) [6–7]. US SRI overcomes this limitation by integrating CEUS, ultrafast imaging, and advanced post-processing, enabling non-invasive visualization of microcirculation and quantification of intratumoral and peritumoral microvascular parameters (e.g., density, directional variance, perfusion index) [8–9]. In contrast to the macroscopic perfusion information from CEUS, US SRI provides complementary microscopic information. Machine

learning is well suited for analyzing such complex multimodal data and has been widely used for disease prediction and stratification [10].

Several studies have confirmed the potential of US SRI for evaluating breast lesions [11–12]. However, most have focused on US SRI as a standalone technique, without assessing its incremental value over conventional ultrasound and CEUS. Moreover, prior research has been limited to intratumoral analysis, ignoring the peritumoral region, which is rich in biomarkers and therapeutic targets [13]. The clinical applicability of existing models also remains unvalidated, especially for complex inflammatory lesions with overlapping imaging features. The integration of macroperfusion (CEUS) and microvascular (US SRI) information may offer novel solutions.

Thus, this study integrates conventional ultrasound, CEUS, and US SRI to extract intrinsic and multi-distance peritumoral features of breast lesions and to build an interpretable machine learning model. Using retrospective modeling, prospective temporally independent validation, and head-to-head comparisons with ultrasound physicians, we will systematically evaluate the relative contribution of each modality, determine the incremental value of US SRI, and verify the model's generalizability and clinical utility. The goal is to provide a reliable auxiliary tool for precise breast lesion diagnosis.

Methods

Research Design and Subjects

This study comprises two distinct phases: a retrospective model development phase (April 2025–January 2026) and a prospective model validation phase (February 2026–March 2026). The study was approved by the institutional ethics committee in accordance with the Declaration of Helsinki (approval number: 2025-LLYJ-001; the date of approval: Jan 14, 2025). Informed consent was waived for retrospective cases, while written informed consent was obtained from all prospective participants.

Retrospective phase inclusion criteria: (1) conventional ultrasound, CEUS, and US SRI completed within one month; (2) no prior treatment or invasive procedures before imaging examinations; (3) complete clinical data, histopathology, and immunohistochemistry results. Exclusion criterion: poor image quality. A total of 210 cases (106 benign, 104 malignant) were included, with histopathology as the gold standard.

Prospective phase used the same inclusion and exclusion criteria. Sample size was estimated based on the retrospective model performance, assuming a validation AUC ≥ 0.95, $\alpha = 0.05$ (two-sided), power = 0.80, using Obuchowski's method [14]. With a benign-to-malignant ratio of approximately 1:2, 32 malignant cases (total = 48) were required. Accounting for a 10% dropout rate, 50–60 patients were planned. Ultimately, 53 cases (18 benign, 35 malignant) were enrolled, meeting the pre-specified sample size. (Fig. 1)

Ultrasound Examination

All ultrasound exams used a Mindray Resona A20 system (5–18 MHz transducer). Patients lay supine with both breasts and axillae fully exposed. Scanning included the lesion's longest-axis plane, its orthogonal plane, and a color Doppler plane.

For CEUS, we selected the plane with the strongest color Doppler flow signal. A 4.8 mL bolus of SonoVue® (sulfur hexafluoride microbubbles) was injected via the left antecubital vein over ≤ 3 seconds, followed immediately by a 5 mL saline flush. Real-time low-MI imaging (0.08–0.12) captured ≥ 120 seconds of dynamic enhancement; full clips were saved.

After ≥ 5 minutes of microbubble washout, a second 4.8 mL SonoVue® bolus was administered. Upon intratumoral microbubble appearance, the probe was held steady, the patient held their breath, and US SRI was triggered. At a 500 Hz frame rate, 6 seconds of contrast data were acquired and processed in real time by the system's built-in US SRI software to generate related features and images.

Feature Extraction

A total of 25 conventional ultrasound and CEUS features were collected, covering lesion morphology, margin, echogenicity, composition, calcification, aspect ratio, posterior acoustic features, size parameters (maximum and minimum diameters, and their ratio), CEUS enhancement patterns (arterial/venous phase, homogeneity, extent, morphology, margin, peripheral radial enhancement, and trending), CDFI grading, BI-RADS categories (on conventional ultrasound and CEUS, as well as post-contrast change), and abnormal lymph nodes. Patient age and anatomical location were also recorded as key clinical parameters.

During the retrospective phase, data were collected blindly by two breast ultrasound physicians (≥ 10 years of experience). Continuous variables were averaged; categorical disagreements were resolved by joint review with a third senior physician. During the prospective phase, data were collected by two different breast ultrasound physicians (≥ 10 years of experience) following strict, standardized clinical protocols.

US SRI parameters were derived from 6-second dynamic sequences. The intratumoral region was manually contoured, and six concentric peritumoral zones (0.5–3.0 mm) were automatically delineated. Twenty-five intratumoral metrics were quantified, including vascular architecture, perfusion, intensity, velocity, directional variance, microvessel diameter, vascular area, and normalized ratios.

In both phases, a single blinded researcher performed image delineation. For inter-rater reliability, 20 cases were randomly selected from each phase and independently delineated by a second researcher (different researchers for the retrospective and prospective phases). ICCs were calculated for all rater pairs.

Feature Selection and Model Development

The analysis was performed using Python (version 3.10.4). The “minimum intensity” feature was excluded (all values were zero). Data were randomly split into training (n = 147) and independent test (n = 63) sets at a 7:3 ratio. In the training set, Lasso regression with 10-fold CV was used for feature selection. Six supervised models (LR, Logistic Regression; DT, Decision Tree; RF, Random Forest; XGBoost, Extreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine; SVM, Support Vector Machine) were developed; each underwent hyperparameter tuning via grid search nested in 10-fold CV. The final models were evaluated on the independent test set to assess predictive performance and generalizability.

Prospective Validation and Comparative Evaluation by Ultrasound Physicians

The best-performing model from the retrospective phase was selected for prospective temporally independent validation. Under blinded conditions, the prospective cohort data were fed into the model to obtain predicted probabilities, which were compared against histopathological diagnoses. Accuracy, sensitivity, specificity, and AUC were calculated to assess generalizability and clinical utility.

One senior (> 10 years) and two junior (< 3 years) ultrasound physicians independently interpreted conventional and CEUS images under blinded conditions (patient age was disclosed). They manually delineated tumor contours on US SRI dynamic sequences to extract microvascular parameters. Integrating all imaging information, each physician rendered a binary (benign/malignant) diagnosis for every lesion.

Statistical Analysis

Statistical analyses were performed using Python (version 3.10.4) and SPSS (version 24.0). Inter-observer agreement for US SRI parameters was assessed using ICC (good reliability defined as ICC > 0.75). The six machine learning models were evaluated for discrimination (AUC with 95% confidence interval/CI), calibration (calibration plots, slope, and Hosmer–Lemeshow test), DCA, and conventional metrics (sensitivity, specificity, accuracy, precision, and F1 score). The best retrospective model was prospectively validated in an independent cohort. SHAP provided feature attribution. Diagnostic performance of three ultrasound physicians was compared against histopathology using the DeLong test (for AUC differences). All tests were two-sided, with $P < 0.05$ considered significant.

Results

Clinical Characteristics

Baseline demographic and clinical characteristics of the overall study population are summarized in Table 1.

In the retrospective cohort (210 patients), the most common benign lesions were fibroadenoma (n = 42), followed by adenosis (n = 22), inflammation (n = 22), and other benign types including benign phyllodes tumor, intraductal papilloma, pseudoangiomatous stromal hyperplasia, and ductal epithelial hyperplasia (n = 20). Among malignant lesions, invasive carcinoma was predominant (n = 93), with fewer cases of ductal carcinoma in situ (DCIS; n = 4) and other malignant types such as lymphoma, neuroendocrine carcinoma, and malignant phyllodes tumor (n = 7).

In the prospective cohort (53 patients), benign lesions comprised fibroadenoma (n = 9), adenosis (n = 2), inflammation (n = 3), and other benign types (n = 4). Malignant lesions included invasive carcinoma (n = 26), DCIS (n = 6), and other malignant types (n = 3).

Table 1
Baseline Characteristics of the Study Population.

Characteristic	Retrospective Cohort				Prospective Cohort	
	Overall (n = 210)	Benign (n = 106)	Malignant (n = 104)	*P* [1]	Overall (n = 53)	*P* [2]
Age (years)	45.76 ± 14.49	38.09 ± 11.09	53.58 ± 13.36	0.280	52.42 ± 12.62	0.002
Tumor location				0.166		0.539
Left breast	96 (45.71%)	43 (40.57%)	53 (50.96%)		27 (50.94%)	
Right breast	114 (54.29%)	63 (59.43%)	51 (49.04%)		26 (49.06%)	
Lesion multiplicity,				0.121		0.571
Solitary	168 (80.00%)	80 (75.47%)	88 (84.62%)		40 (75.47%)	
Multiple	42 (20.00%)	26 (24.53%)	16 (15.38%)		13 (24.53%)	
Clinical presentation				<0.001		0.114
Palpable mass	117 (55.72%)	41 (38.68%)	76 (73.08%)		39 (73.58%)	
Nipple discharge	5 (2.38%)	4 (3.77%)	1 (0.96%)		0 (0.00%)	
Inflammatory symptoms [3]	15 (7.14%)	15 (14.15%)	0 (0.00%)		3 (5.66%)	
Incidentally found	73 (34.76%)	46 (43.40%)	27 (25.96%)		11 (20.76%)	

Note: Data are mean ± SD or n (%). Continuous: t-test or Mann-Whitney U test; categorical: χ^2 test or Fisher's exact test. [1] Benign vs. malignant in retrospective cohort. [2] Retrospective vs. prospective cohorts. [3] Inflammatory symptoms: redness, swelling, pain. Abbreviations: SD, standard deviation.

Consistency Assessment

Twenty cases were randomly selected from each cohort for independent blinded re-delineation of US SRI parameters by different researchers. ICCs for all parameters were > 0.75 (retrospective: 0.78–0.95; prospective: 0.80–0.94), indicating good to excellent inter-observer reliability.

Retrospective Model Development

Lasso regression with 10-fold CV identified eight key predictors independently associated with benign and malignant breast lesions from the original feature set. Compared with the benign group, the malignant group showed significantly higher age, minimum diameter, BI-RADS on CEUS, and 1.0pVD, but a significantly lower ratio of diameter ($P < 0.05$). Dir.Var, 0.5pPI, and anatomical location did not differ significantly between groups ($P > 0.05$). (Table 2; Fig. 2)

Table 2
Distribution of Selected Predictors in the Training Set.

Characteristic	Benign (n = 74)	Malignant (n = 73)	Statistic	P-value
Age (years)	36.8 ± 11.2	52.1 ± 13.7	U = 1019.5	< 0.001
Minimum diameter (mm)	9.9 ± 4.9	15.1 ± 6.6	U = 1282.0	< 0.001
Ratio of diameter	2.2 ± 0.9	1.8 ± 0.5	U = 3621.5	< 0.001
Dir.Var	3687.4 ± 983.6	3957.2 ± 3655.2	U = 3019.0	0.219
0.5pPI	9.2 ± 4.6	9.7 ± 5.3	U = 2595.5	0.684
1.0pVD	28.0 ± 12.0	32.5 ± 12.4	U = 2123.5	0.025
Anatomical location			$\chi^2 = 1.42$	0.840
Upper outer quadrant	30 (40.5)	34 (46.6)		
Lower outer quadrant	17 (23.0)	12 (16.4)		
Upper inner quadrant	16 (21.6)	18 (24.7)		
Lower inner quadrant	6 (8.1)	5 (6.8)		
Central region	5 (6.8)	4 (5.5)		
BI-RADS on CEUS			$\chi^2 = 115.19$	< 0.001
Category 2/3	47 (63.5)	4 (5.5)		
Category 4a	21 (28.4)	0 (0.0)		
Category 4b	5 (6.8)	10 (13.7)		
Category 4c	1 (1.4)	14 (19.2)		
Category 5	0 (0.0)	45 (61.6)		

Note: Data are mean ± SD or n (%). Continuous: t-test or Mann-Whitney U test; categorical: χ^2 test or Fisher's exact test. Abbreviations: SD, standard deviation; Dir.Var, direction variance; 0.5pPI, perfusion index at 0.5 mm peritumoral region; 1.0pVD, vessel density at 1.0 mm peritumoral region; CEUS, contrast-enhanced ultrasound.

Six machine learning models were built based on the above characteristics. Their internal stability and performance were assessed using 10-fold CV on the training set; mean AUCs (0.9685) and standard

deviations (0.03–0.045) are shown in Table 3. Ensemble models (XGBoost, LightGBM, RF) outperformed individual models (LR, DT, SVM).

Table 3
10-fold Cross-validation AUC of Six Machine Learning Models.

Model	10-fold CV AUC (mean ± SD)	95% Confidence Interval
LR	0.9543 ± 0.0423	[0.8714, 1.0372]
DT	0.9648 ± 0.0387	[0.8889, 1.0407]
RF	0.9663 ± 0.0439	[0.8803, 1.0523]
XGBoost	0.9742 ± 0.0317	[0.9121, 1.0363]
LightGBM	0.9756 ± 0.0301	[0.9166, 1.0346]
SVM	0.9755 ± 0.0342	[0.9085, 1.0425]

Note: Confidence intervals were calculated using the normal approximation method; some upper bounds exceed 1.0, which is a known phenomenon when AUC is close to 1 and sample size is limited. Abbreviations: AUC, area under the curve; SD, standard deviation; LR, Logistic Regression; DT, Decision Tree; RF, Random Forest; XGBoost, Extreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine; SVM, Support Vector Machine.

Predictive performance for all six models is summarized in Table 4 and Fig. 3 (training and independent test sets). In the training set, AUCs ranged from 0.9455 to 1.0000; ensemble models (XGBoost, LightGBM, RF) outperformed single-learner models (LR, DT, SVM). In the independent test set, AUCs ranged from 0.8896 to 0.9909—again with ensembles superior. The RF model achieved the highest test-set AUC (0.9909; 95% CI: 0.9675–1.0143), indicating strong generalizability.

Table 4
Predictive Performance of six Machine Learning Models.

Dataset	Model	AUC (95% CI)	Sensitivity	Specificity	Accuracy	Precision	F1 Score
Training	LR	0.9455 (0.9088, 0.9822)	0.9315	0.9595	0.9456	0.9577	0.9444
	DT	0.9864 (0.9677, 1.0051)	0.9452	0.9189	0.9320	0.9200	0.9324
	RF	0.9957 (0.9852, 1.0063)	0.9452	0.9730	0.9592	0.9718	0.9583
	XGBoost	1 (1,1)	1	1	1	1	1
	LightGBM	0.9884 (0.9711, 1.0057)	0.9452	0.9730	0.9592	0.9718	0.9583
	SVM	0.9787 (0.9554, 1.0020)	0.8904	0.9730	0.9320	0.9701	0.9286
Test	LR	0.8896 (0.8122, 0.9670)	0.9355	0.8438	0.8889	0.8529	0.8923
	DT	0.8931 (0.8169, 0.9694)	0.9677	0.7500	0.8571	0.7895	0.8696
	RF	0.9909 (0.9675, 1.0143)	0.9677	0.8438	0.9048	0.8571	0.9091
	XGBoost	0.9748 (0.9361, 1.0135)	0.9355	0.8750	0.9048	0.8788	0.9063
	LightGBM	0.9803 (0.9461, 1.0146)	0.9677	0.8750	0.9206	0.8824	0.9231
	SVM	0.9818 (0.9489, 1.0148)	0.9355	0.8750	0.9048	0.8788	0.9063

Note: LR, Logistic Regression; DT, Decision Tree; RF, Random Forest; XGBoost, Extreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine; SVM, Support Vector Machine; CI, Confidence Interval.

Calibration performance of the six models in the test set was assessed using calibration curves and quantitative metrics, reported in Table 5 and Fig. 3. The RF model had the steepest calibration slope

(2.5205), suggesting mild underfitting, but its Brier score (0.0655) was second lowest—only surpassed by SVM. Its Eavg (0.1453) indicates mid-probability miscalibration, yet overall calibration remained excellent. Clinically, mild underfitting—yielding conservative probability estimates—is preferable to overfitting, which risks overconfidence in borderline cases.

Table 5
Calibration Metrics of six Machine Learning Models on the Test Set.

Model	Brier score	Eavg	Calibration slope	Calibration intercept
LR	0.0791	0.0978	0.7867	−1.3021
DT	0.1170	0.1517	0.1796	−10.3823
RF	0.0655	0.1453	2.5205	−0.5395
XGBoost	0.0798	0.0782	0.6468	−1.7447
LightGBM	0.0699	0.1209	1.8269	−0.5506
SVM	0.0580	0.0807	1.2282	−1.2408

Note: LR, Logistic Regression; DT, Decision Tree; RF, Random Forest; XGBoost, Extreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine; SVM, Support Vector Machine.

DCA assessed the clinical net benefit of all six models in the test set (Fig. 3). Across threshold probabilities, all models outperformed both “treat-all” and “treat-none” strategies, supporting clinical utility. The “treat-all” strategy represents the net benefit if all patients were referred for biopsy (i.e., treat everyone as malignant), whereas the “treat-none” strategy represents the net benefit if no biopsy were performed (i.e., treat everyone as benign). The net benefit was calculated as described by Vickers et al [15]. The XGBoost model had the widest applicable threshold range (0.01–0.99), followed by SVM model (0.06–0.99) and RF model (0.07–0.98). At a 0.5 threshold, XGBoost, RF, and SVM avoided 42.2, 42.1, and 42.2 unnecessary biopsies per 100 patients, respectively.

Comparative evaluation of six machine learning models on the training and test sets, showing ROC curves (A1, B1), calibration curves (A2, B2), decision curves (A3, B3), and radar charts of key metrics (A4, B4).

On balance, RF exhibited the best discriminative power (achieving the highest AUC on the test set), SVM achieved optimal calibration performance, and XGBoost demonstrated the greatest net clinical benefit. Given clinical priorities—generalizability and stability—the RF model was selected for prospective temporally independent validation and comparative assessment against radiologists.

Model Explanation

To interpret the RF model—the top-performing model—we applied SHAP analysis to quantify and visualize feature contributions. Figure 4A shows mean absolute SHAP values, ranking feature importance. BI-RADS on CEUS was the strongest predictor, far exceeding all others—highlighting its

central role in differentiating benign from malignant breast lesions. Age and minimum diameter ranked second and third, confirming their substantial diagnostic value. Ratio of diameter, Dir.Var, 1.0pVD, and 0.5pPI showed similar, moderate importance. Anatomical location contributed negligibly, indicating minimal influence on predictions.

Figure 4B (SHAP beeswarm plot) reveals how each feature affects predictions across its value range. Higher BI-RADS on CEUS, age, and minimum diameter consistently yield positive SHAP values—indicating increased malignancy risk. A high ratio of diameter (mostly red, clustered in the positive region) suggests malignancy when > 1. In contrast, Dir.Var, 1.0pVD, and 0.5pPI show complex, nonlinear patterns—implying their effects depend on interactions with other features.

Prospective Temporally Independent Validation

The retrospective top-performing RF model was validated prospectively. It achieved an AUC of 0.9492 (95% CI: 0.8945–1.0000), accuracy of 92.45%, sensitivity of 97.14%, and specificity of 83.33%. Though its AUC declined slightly from 0.9909 (retrospective test set) to 0.9492—a drop of 0.0417—it remains excellent, confirming robust generalizability.

Compared with junior ultrasound physicians (AUC-1: 0.8175, 95% CI: 0.7135–0.9215, $P = 0.0309$; AUC-2: 0.8032, 95% CI: 0.6961–0.9102, $P = 0.0193$), the RF model performed significantly better. Its performance was comparable to that of the senior physician (AUC: 0.9286, 95% CI: 0.8592–0.9979, $P = 0.6571$). (Fig. 5)

Discussion

Unlike prior studies—those relying solely on conventional ultrasound and/or CEUS [16–18], or those validating only US SRI [12, 19]—this study integrates macroscopic CEUS (reflecting overall tumor perfusion) with microscopic US SRI (revealing intratumoral microvascular architecture). This dual-scale approach captures both tumor-level biological behavior and microenvironmental heterogeneity, delivering more comprehensive imaging evidence for accurate breast lesion diagnosis.

To avoid overfitting and underfitting, six machine learning models were built and evaluated using 10-fold CV, independent testing, and prospective temporally independent validation. All models achieved mean 10-fold CV AUC > 0.95 (SD: 0.03–0.045). In prospective temporally independent validation, the RF model's AUC decreased by only 0.04 compared with the retrospective test set, demonstrating excellent generalizability and low risk of overfitting. This multi-model, multi-level validation paradigm enhances the reliability and clinical translatability of our findings.

SHAP analysis identified BI-RADS on CEUS, age, and minimum diameter as the top three drivers of the RF model's predictions, with BI-RADS on CEUS ranking first. This reflects the model's alignment with clinical expertise: BI-RADS integrates morphology, margin, and enhancement pattern—key criteria radiologists use to assess malignancy [20]. Age and minimum diameter were significantly higher in

malignant versus benign lesions ($P < 0.05$), consistent with established knowledge: age is a known breast cancer risk factor, and size underpins TNM staging [21]. Ratio of diameter > 1—indicating non-circular morphology—was strongly associated with positive SHAP values, supporting its role as a malignancy marker, consistent with BI-RADS guidance on shape assessment [20]. In contrast, anatomical location contributed negligibly to predictions, aligning with clinical reality: breast cancer incidence does not vary meaningfully by quadrant, and location is not an independent risk factor [22–23]. Collectively, the model's SHAP-derived interpretations closely mirror clinical prior knowledge, strengthening clinician trust and supporting clinical translatability.

SHAP analysis shows Dir.Var, 1.0pVD, and 0.5pPI contribute minimally to predictions (mean |SHAP| < 0.02)—far less than core clinical features—yet all were retained by Lasso regression, confirming their non-redundant predictive value.

The SHAP beeswarm plot reveals that high Dir.Var values cluster in the positive SHAP region, linking increased intratumoral vessel orientation heterogeneity to higher malignancy risk—consistent with the disorganized, tortuous microvasculature typical of malignant tumors [24]; that the influence direction of 0.5pPI exhibits a nonlinear distribution, where both low and high values may indicate malignancy—likely reflecting coexisting hypoperfused (e.g., necrotic) and hyperperfused (e.g., angiogenic) peritumoral regions across pathological subtypes [25]; and that, consistent with the established association between vascular density and tumor-induced angiogenesis [26], a high value of 1.0pVD predicts malignancy, suggesting that such angiogenesis also occurs in the peritumoral region.

The limited contribution of US SRI parameters likely stems from three factors: (i) information overlap—peritumoral microvascular changes are partially captured by CEUS enhancement patterns; (ii) sample size—our retrospective cohort (n = 201) may be underpowered to robustly estimate their independent effects; and (iii) biological heterogeneity—not all malignant lesions exhibit prominent peritumoral microvascular alterations, and manifestations vary across molecular and pathological subtypes, diluting overall signal.

Limited contribution does not imply lack of value. First, these parameters are retained in the model and show complex nonlinear patterns in the SHAP beeswarm plot—indicating they capture unique microvascular information beyond core clinical features and are non-redundant. Second, they provide new imaging evidence supporting the hypothesis that "the peritumoral microenvironment plays a key role in breast cancer progression" [13]. Third, in clinical practice, they offer critical auxiliary information when core features are ambiguous—e.g., differentiating lesions with overlapping imaging features, particularly complex inflammatory lesions from malignancies—helping reduce unnecessary needle biopsies.

Notably, Lasso regression retained only 1.0pVD and 0.5pPI—both from the 0.5–1.0 mm peritumoral zone—while excluding all parameters at longer distances (1.5–3.0 mm). This suggests the tumor-adjacent region (0.5–1.0 mm) is the most information-rich for detecting tumor-associated microvascular changes; beyond 1.5 mm, signals likely reflect normal tissue and dilute tumor-specific alterations. By

screening multi-distance US SRI parameters in parallel, this study identifies the 0.5–1.0 mm zone as the information-enriched zone and the ≥ 1.5 mm zone as information-redundant—providing a practical basis for optimizing peritumoral region definition and reducing feature dimensionality in future work.

DCA shows that RF, XGBoost, and SVM yield higher net benefits than “treat-all” or “treat-none” across a broad range of threshold probabilities. At a 0.5 threshold, they avoid 42.2, 42.1, and 42.2 unnecessary biopsies per 100 patients, respectively—demonstrating strong clinical utility. Among them, RF delivers consistently high net benefit in the clinically relevant 0.2–0.6 threshold range. Combined with its superior discrimination, calibration, and generalization, this confirms RF as the optimal model for routine diagnosis.

In the prospective cohort, the RF model’s AUC (0.9492) significantly exceeded junior physicians’ (0.8032–0.8175; $P < 0.05$) and matched senior physicians’ (0.9286; $P = 0.6571$). This shows that the model can boost junior physicians’ diagnostic accuracy and serve as a reliable second opinion for senior physicians—reducing misdiagnosis and missed diagnosis in challenging cases or during fatigue. Clinical adoption of the model depends not only on accuracy but also on physicians’ trust in its decision-making [27]. SHAP makes the model’s “black-box” decisions interpretable by visualizing feature contributions. Hur et al. [28] showed that explanatory information significantly improves clinicians’ acceptance of artificial intelligence systems, with trust, satisfaction, and usability positively correlated with acceptance, confirming the value of interpretability in building clinical trust.

This study has several limitations. First, both retrospective and prospective cohorts were single-center, making the prospective temporally independent validation not truly external; multi-center validation is needed. Second, there is no consensus on US SRI parameter stability or peritumoral zone definition; our 0.5–1.0 mm finding requires external verification. Third, SHAP reveals associations, not causation, and the RF model’s calibration showed under-fitting. Fourth, the radiologist comparison involved only three readers and retrospective review, limiting real-world generalizability. Future work will pursue multi-center external validation, integrate pathomics to explore biological mechanisms, apply probability calibration (e.g., Platt scaling), and conduct larger-scale prospective reader studies.

Conclusions

This study integrated multimodal ultrasound features and built a robust, generalizable machine learning model for breast lesion diagnosis—using six algorithms, 10-fold CV, and prospective temporally independent validation to rigorously control overfitting. The RF model achieved excellent discrimination and generalization: AUC = 0.9909 (retrospective test set) and 0.9492 (prospective cohort). SHAP analysis confirms core clinical features as dominant, quantifies the incremental value of US SRI, and validates the complementary utility of macro- and micro-scale imaging information. Compared with ultrasound physicians, the model boosts junior physicians’ diagnostic accuracy and provides seniors with reliable second opinions—reducing unnecessary biopsies. Overall, this work advances multimodal ultrasound radiomics for precise breast lesion diagnosis and accelerates its clinical translation.

Abbreviations

AUC, area under the curve;

BI-RADS, breast imaging reporting and data system;

CDFI, color Doppler flow imaging;

CEUS, contrast-enhanced ultrasound;

CI, confidence interval;

CV, cross-validation;

DCA, decision curve analysis;

DCIS, ductal carcinoma in situ;

Dir.Var, directional variance;

DT, Decision Tree;

ICC, Intraclass correlation coefficient;

LightGBM, Light Gradient Boosting Machine;

LR, Logistic Regression;

RF, Random Forest;

SD, standard deviation;

SHAP, SHapley Additive exPlanations;

SVM, Support Vector Machine;

US SRI, ultrasound super-resolution imaging;

XGBoost, Extreme Gradient Boosting;

0.5pPI, 0.5mm peritumoral perfusion index;

1.0pVD, 1.0mm peritumoral vascular density.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of the First Affiliated Hospital of the University of Science and Technology of China in accordance with the Declaration of Helsinki (approval number: 2025-LLYJ-001; the date of approval: Jan 14, 2025). Informed consent was waived for retrospective cases, while written informed consent was obtained from all prospective participants.

Consent for publication

Figure 2 uses images of two patients as examples. Consent for publication has been obtained from the corresponding patients, and we can provide copies for review at any time.

Availability of data and materials

The data supporting the findings of this study are not publicly available due to patient confidentiality and institutional data governance policies. De-identified data may be available from the corresponding author upon reasonable request, subject to institutional approval from the Institutional Review Board of the First Affiliated Hospital of the University of Science and Technology of China (Approval No.: 2025-LLYJ-001).

Competing interests

The authors declare that they have no competing interests.

Funding

The study was supported by a research grant of Kun Tao from the 2025 Institutional Youth Fund of Anhui Provincial Cancer Hospital (No. 2025YJQN006).

Authors' contributions

Bin Shi: Conceptualization, Methodology, Writing—Original draft preparation; Yang Yang: Ultrasound data curation, Writing—Original draft preparation; Jing Zhang: Investigation, Ultrasound data curation; Yun Yuan: Clinical data curation; Chao Wei: Methodology; Kun Tao: Conceptualization, Writing—Reviewing and Editing; Lei Ye: Supervision, Writing—Reviewing and Editing.

Acknowledgements

Not Applicable.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63. 10.3322/caac.21834.
2. Beckmann MW, Niederacher D, Schnürch HG, Gusterson BA, Bender HG. Multistep carcinogenesis of breast cancer and tumour heterogeneity. J Mol Med (Berl). 1997;75(6):429–39.

10.1007/s001090050128.

3. Xu H, Xu B. Epidemiology, early detection, and management of breast cancer in China: A comprehensive review. Chin J Cancer Res. 2025;37(6):882–99. 10.21147/j.issn.1000-9604.2025.06.02.
4. Chen Y, Lu J, Li J, Liao J, Huang X, Zhang B. Evaluation of diagnostic efficacy of multimode ultrasound in BI-RADS 4 breast neoplasms and establishment of a predictive model. Front Oncol. 2022;12:1053280. 10.3389/fonc.2022.1053280.
5. Yoo J, Je BK, Choo JY. Ultrasonographic Demonstration of the Tissue Microvasculature in Children: Microvascular Ultrasonography Versus Conventional Color Doppler Ultrasonography. Korean J Radiol. 2020;21(2):146–158. doi: 10.3348/kjr.2019.0500. Erratum in: Korean J Radiol. 2020;21(4):509. 10.3348/kjr.2020.0208
6. Chou CP, Huang JS, Wang JS, Pan HB. Contrast-enhanced ultrasound features of breast capillary hemangioma: a case report and review of literature. J Ultrasound. 2022;25(1):103–6. 10.1007/s40477-020-00550-y.
7. Luan S, Yu X, Lei S, Ma C, Wang X, Xue X, Ding Y, Ma T, Zhu B. Deep learning for fast super-resolution ultrasound microvessel imaging. Phys Med Biol. 2023;68(24). 10.1088/1361-6560/ad0a5a.
8. Christensen-Jeffries K, Couture O, Dayton PA, Eldar YC, Hynynen K, Kiessling F, O'Reilly M, Pinton GF, Schmitz G, Tang MX, Tanter M, van Sloun RJG. Super-resolution Ultrasound Imaging. Ultrasound Med Biol. 2020;46(4):865–91. 10.1016/j.ultrasmedbio.2019.11.013.
9. Bodard S, Denis L, Hingot V, Chavignon A, Hélénon O, Anglicheau D, Couture O, Correas JM. Ultrasound localization microscopy of the human kidney allograft on a clinical ultrasound scanner. Kidney Int. 2023;103(5):930–5. 10.1016/j.kint.2023.01.027.
10. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1216–9. 10.1056/NEJMp1606181.
11. Lei YM, Liu C, Hu HM, Li N, Zhang N, Wang Q, Zeng SE, Ye HR, Zhang G. Combined use of super-resolution ultrasound imaging and shear-wave elastography for differential diagnosis of breast masses. Front Oncol. 2024;14:1497140. 10.3389/fonc.2024.1497140.
12. Zhang G, Lei YM, Li N, Yu J, Jiang XY, Yu MH, Hu HM, Zeng SE, Cui XW, Ye HR. Ultrasound super-resolution imaging for differential diagnosis of breast masses. Front Oncol. 2022;12:1049991. 10.3389/fonc.2022.1049991.
13. Zhang S, Regan K, Najera J, Grinstaff MW, Datta M, Nia HT. The peritumor microenvironment: physics and immunity. Trends Cancer. 2023;9(8):609–23. 10.1016/j.trecan.2023.04.004.
14. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Stat Med. 1997;16(13):1529–42. 10.1002/(sici)1097-0258(19970715)16:13%3C1529::aid-sim565%3E3.0.co;2-h.
15. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565–74. 10.1177/0272989X06295361.

16. Hu Q, Wang XY, Zhu SY, Kang LK, Xiao YJ, Zheng HY. Meta-analysis of contrast-enhanced ultrasound for the differentiation of benign and malignant breast lesions. Acta Radiol. 2015;56(1):25–33. 10.1177/0284185113517115.
17. Du J, Wang L, Wan CF, Hua J, Fang H, Chen J, Li FH. Differentiating benign from malignant solid breast lesions: combined utility of conventional ultrasound and contrast-enhanced ultrasound in comparison with magnetic resonance imaging. Eur J Radiol. 2012;81(12):3890–9. 10.1016/j.ejrad.2012.09.004.
18. Xiao X, Dong L, Jiang Q, Guan X, Wu H, Luo B. Incorporating Contrast-Enhanced Ultrasound into the BI-RADS Scoring System Improves Accuracy in Breast Tumor Diagnosis: A Preliminary Study in China. Ultrasound Med Biol. 2016;42(11):2630–8. 10.1016/j.ultrasmedbio.2016.07.005.
19. Yang L, Ma Z. Nomogram Based on Super-Resolution Ultrasound Images Outperforms in Predicting Benign and Malignant Breast Lesions. Breast Cancer (Dove Med Press). 2023;15:867–78. 10.2147/BCTT.S435510.
20. Portaluri A, Trimarchi R, Wang X, Zhang T, Aricò FM, Condorelli E, Zagami P, Ascenti V, Ieni A, Marino MA. Can the new combined CEUS-BI-RADS score (CB score) reduce the number of breast false positive biopsies? Results from a monocentric study. Radiol Med. 2026. 10.1007/s11547-026-02179-z.
21. Ren W, Chen M, Qiao Y, Zhao F. Global guidelines for breast cancer screening: A systematic review. Breast. 2022;64:85–99. 10.1016/j.breast.2022.04.003.
22. Janni W, Rack B, Sommer H, Schmidt M, Strobl B, Rjosk D, Klanner E, Thieleke W, Gerber B, Friese K, Dimpfl T. Intra-mammary tumor location does not influence prognosis but influences the prevalence of axillary lymph-node metastases. J Cancer Res Clin Oncol. 2003;129(9):503–10. 10.1007/s00432-003-0465-3.
23. Wei Y, Liu P, Liu X, Wang M, Liu D, Cui H, Lin S, Wu H, Ma X, Kang H. Analysis of the Relationship Between Primary Tumor Site and Clinicopathological Characteristics and Survival Prognosis of Breast Cancer Patients Based on SEER Database. Clin Breast Cancer. 2024;24(8):728–45. 10.1016/j.clbc.2024.08.021.
24. Jain RK. Antiangiogenesis strategies revisited: from starving tumors to alleviating hypoxia. Cancer Cell. 2014;26(5):605–22. 10.1016/j.ccell.2014.10.006.
25. Vaupel P, Mayer A, Höckel M. Tumor hypoxia and malignant progression. Methods Enzymol. 2004;381:335–54. 10.1016/S0076-6879(04)81023-1.
26. Park AY, Kwon M, Woo OH, Cho KR, Park EK, Cha SH, Song SE, Lee JH, Cha J, Son GS, Seo BK. A Prospective Study on the Value of Ultrasound Microflow Assessment to Distinguish Malignant from Benign Solid Breast Masses: Association between Ultrasound Parameters and Histologic Microvessel Densities. Korean J Radiol. 2019;20(5):759–72. 10.3348/kjr.2018.0515.
27. Roychowdhury S, Lanfranchi V, Mazumdar S. Evaluating explanation performance for clinical decision support systems for non-imaging data: A systematic literature review. Comput Biol Med. 2025;197(Pt A):110944. 10.1016/j.compbiomed.2025.110944.

28. Hur S, Lee Y, Park J, Jeon YJ, Cho JH, Cho D, Lim D, Hwang W, Cha WC, Yoo J. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. NPJ Digit Med. 2025;8(1):578. 10.1038/s41746-025-01958-8.

Figures

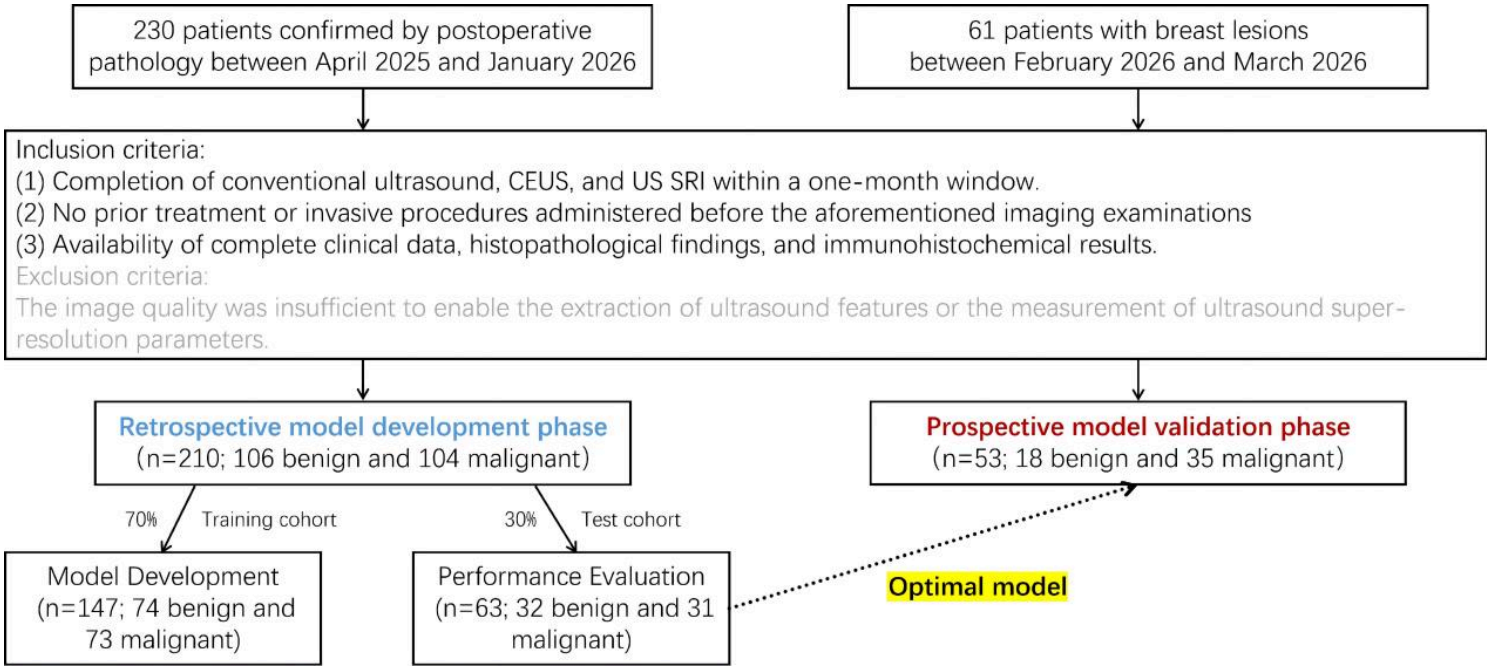


Figure 1

Flow Chart of this Study.

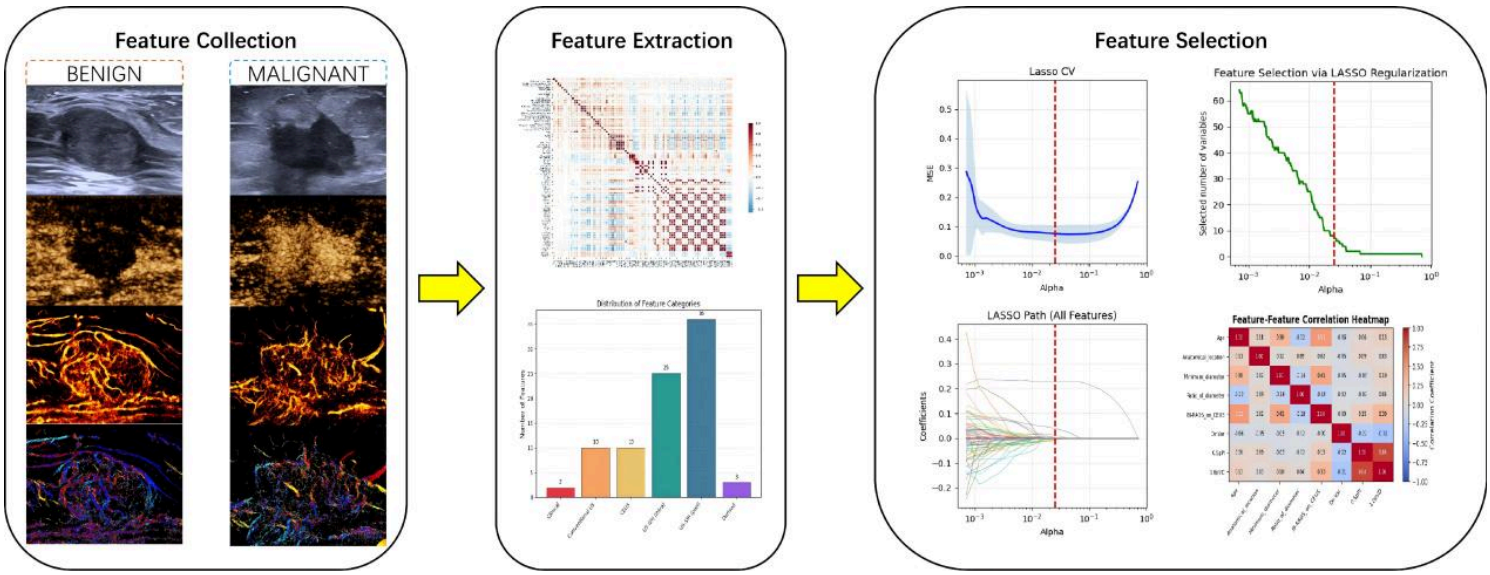


Figure 2

Feature Extraction and Feature Selection.

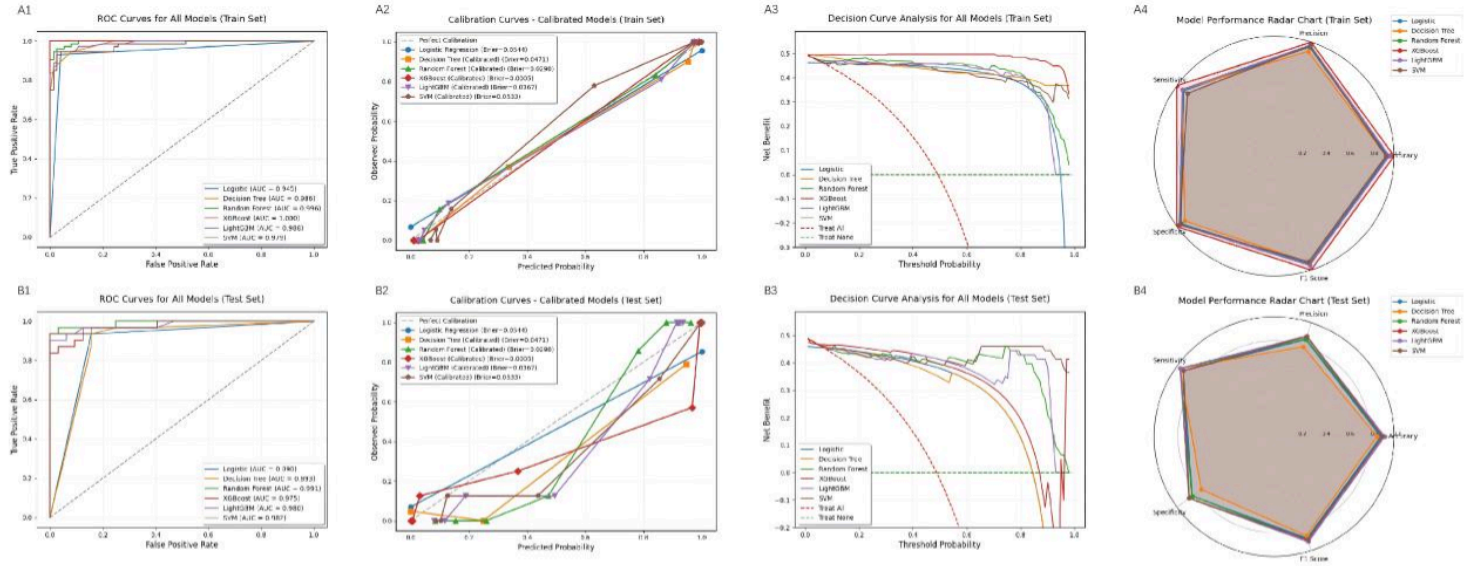


Figure 3

Comparison of Predictive Performance.

Comparative evaluation of six machine learning models on the training and test sets, showing ROC curves (A1, B1), calibration curves (A2, B2), decision curves (A3, B3), and radar charts of key metrics (A4, B4).

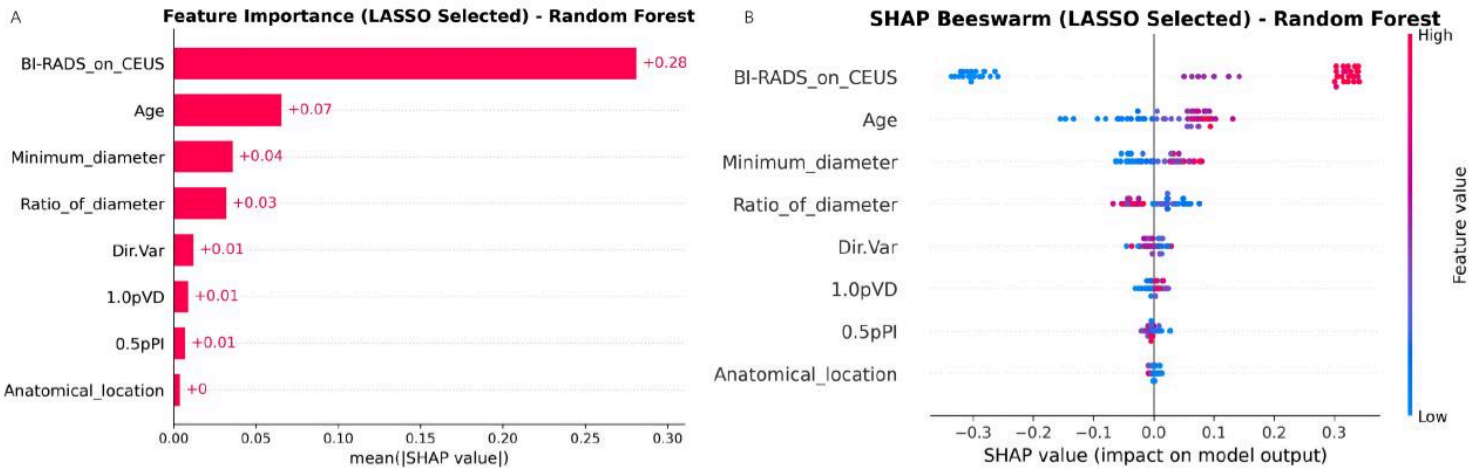


Figure 4

SHAP Interpretation of the RF Model.

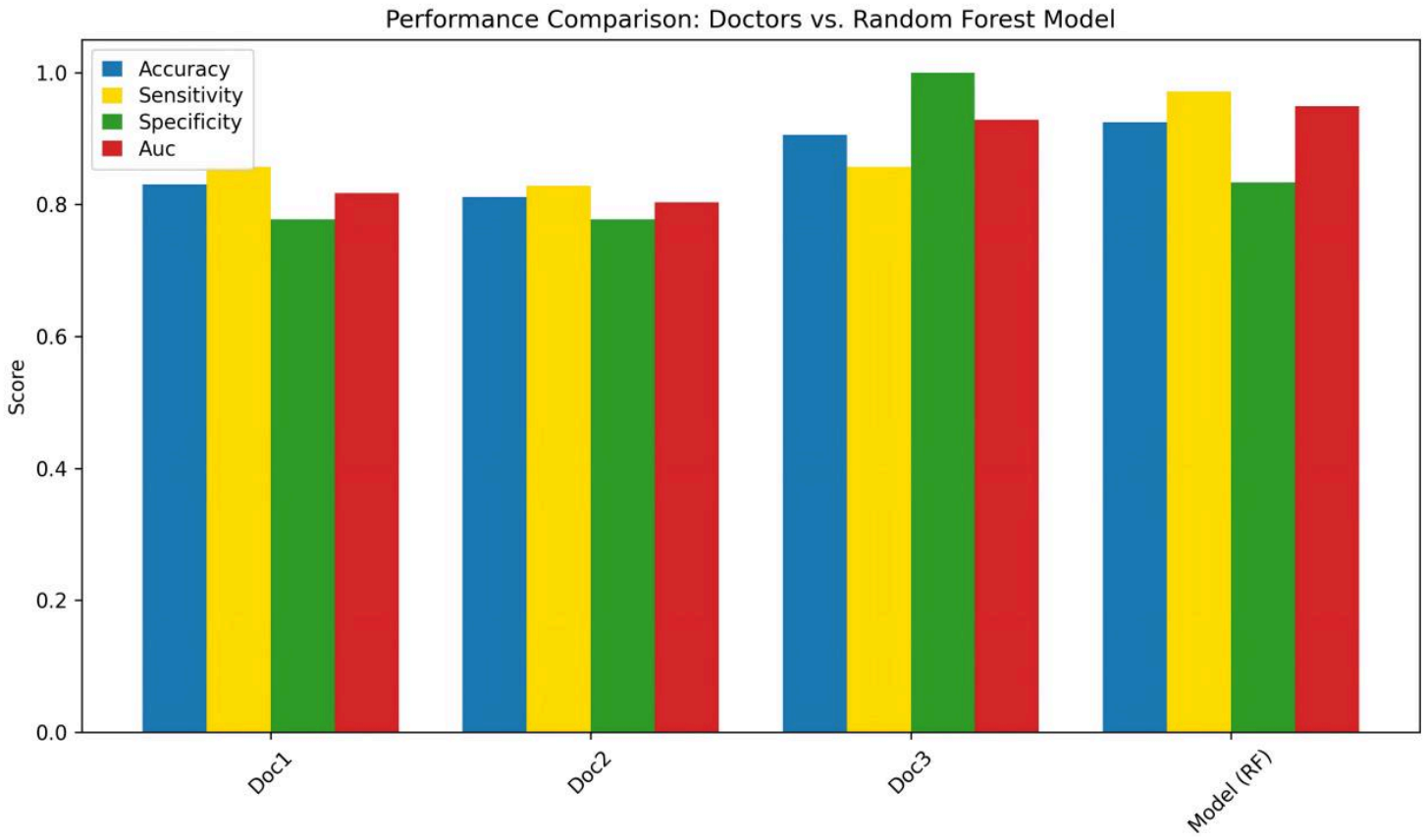


Figure 5

Prospective Temporally Independent Validation Comparative Evaluation by Ultrasound Physicians.