Original Article

Development and validation of a semi-automatic radiomics ensemble model for preoperative evaluation of breast masses in mammotome-assisted minimally invasive resection

Zhenfeng Huang[1#], Qingqing Zhu[2#], Yijie Li[1#], Kunyi Wang[1], Yideng Zhang[1], Qiaowei Zhong[3], Yi Li[1], Qingan Zeng[1], Haihong Zhong[4]

[1]Department of Thyroid & Breast Surgery, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China; [2]Department of Proctology, Wuhan Third Hospital, Tongren Hospital of Wuhan University, Wuhan, China; [3]Zhuhai Campus of Zunyi Medical University, Zhuhai, China; [4]Department of Radiology, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China

Contributions: (I) Conception and design: Z Huang, Q Zhu; (II) Administrative support: H Zhong, Q Zeng; (III) Provision of study materials or patients: Yijie Li; (IV) Collection and assembly of data: K Wang, Y Zhang, Q Zhong; (V) Data analysis and interpretation: Z Huang, Yi Li; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Haihong Zhong, BS. Department of Radiology, The Fifth Affiliated Hospital, Sun Yat-sen University, No. 52 Meihua East Road, Zhuhai 519000, China. Email: zhonghh7@mail.sysu.edu.cn; Qingan Zeng, MS; Yi Li, MD. Department of Thyroid & Breast Surgery, The Fifth Affiliated Hospital, Sun Yat-sen University, No. 52 Meihua East Road, Zhuhai 519000, China. Email: Zengqa123@163.com; liyi8@sysu.edu.cn.

Background: Accurate preoperative differentiation of breast masses is critical for guiding individualized treatment strategies in Mammotome-assisted minimally invasive resection. While radiomics shows promise, existing methods rely on manual delineation, which is time-consuming and subjective. This study developed an ultrasound-based semi-automatic segmentation ensemble model to improve preoperative assessment.

Methods: We retrospectively analyzed preoperative ultrasound images from 773 patients (543 tumors, 230 non-tumors). Semi-automatic segmentation was performed using DeepLabv3_ResNet50 and fully convolutional network (FCN)_ResNet50. Radiomic and deep transfer learning (DTL) features were extracted to construct radiomic, deep learning, and combined models. An ensemble strategy integrated these with clinical models. Performance was evaluated via receiver operating characteristic (ROC) curves and decision curve analysis (DCA).

Results: The cohort included 543 tumor patients and 230 non-tumor patients (95 adenosis, 135 other benign lesions). The semi-automatic segmentation model, DeepLabv3_ResNet50, achieved a peak global accuracy of 99.4% and an average Dice coefficient of 92.0% at its best epoch. On the other hand, the FCN_ResNet50 model exhibited a peak global accuracy of 99.5% and an average Dice coefficient of 93.7% at its best epoch. In the task of predicting tumor and non-tumor patients, age, maximum diameter, and BI-RADS (Breast Imaging Reporting and Data System) classification were ultimately identified as key indicators, and the stacking model ultimately demonstrated an area under the curve (AUC) of 0.890 in the training cohort (with a sensitivity of 0.844 and a specificity of 0.815) and an AUC of 0.780 in the testing cohort (with a sensitivity of 0.713 and a specificity of 0.739). In the task of predicting adenosis and other lesion types, focus emerged as a crucial factor, and the stacking model achieved an AUC of 0.813 in the training cohort (with a sensitivity of 0.613 and a specificity of 0.859) and an AUC of 0.771 in the testing cohort (with a sensitivity of 0.759 and a specificity of 0.765).

Conclusions: Our study has established an ensemble learning model grounded in semi-automatic segmentation techniques. This model accurately distinguishes between tumor and non-tumor patients preoperatively, as well as discriminating adenosis from other lesion types among the non-tumor cohort, thus providing valuable insights for individualized treatment planning. The proposed stacking model demonstrates significant clinical utility by reducing unnecessary biopsies and saving diagnostic time compared to manual review. These improvements directly address the challenges of overtreatment and

diagnostic delays in breast lesion management. By enhancing preoperative accuracy, our model supports tailored surgical planning and alleviates patient anxiety associated with indeterminate diagnoses.

Keywords: Breast lesion; ultrasound (US); mammotome; radiomics; deep learning

Submitted Oct 11, 2024. Accepted for publication Mar 04, 2025. Published online Mar 26, 2025.
doi: 10.21037/gs-24-440
View this article at: https://dx.doi.org/10.21037/gs-24-440

Introduction

According to the statistics of the International Agency for Research on Cancer, breast cancer is one of the most common malignancies globally, with a continuously high incidence and mortality rate, particularly among women (1,2). The significantly higher incidence of breast cancer compared to other malignancies is an indisputable fact (3). Due to its high incidence and mortality rate, early screening and diagnosis of breast cancer have become of paramount importance (4). Ultrasound (US) has played a significant role in breast cancer screening, enabling doctors to detect and evaluate abnormalities in the breast, including lumps, cystic lesions, and intraluminal proliferations of the mammary ducts (5,6).

In recent years, with the continuous advancement of minimally invasive breast techniques and the increasing aesthetic demands of patients, the US-guided Mammotome system has been widely used for accurate biopsy of suspicious lesions or removal of benign breast lesion (7). The Mammotome minimally invasive excisional biopsy is commonly employed for the treatment and diagnosis of benign breast lesions with a diameter of less than 2 cm, particularly those that cannot be definitively diagnosed by conventional means. Preoperative assessment of the nature of breast lesion is crucial; by improving diagnostic accuracy, physicians can gain a better understanding of the biological characteristics of the lesion, thereby developing personalized treatment plans that enhance efficacy and reduce unnecessary surgical interventions. Moreover, this assessment helps to better determine the suitability of the Mammotome minimally invasive excisional biopsy, ensuring surgical success rates and providing comprehensive information for postoperative follow-up and management. For patients, making informed surgical choices, especially when distinguishing between benign and malignant lesion, significantly impacts their psychological well-being and treatment expectations. Therefore, preoperative assessment is not only a technical consideration but also a vital component of overall patient management (8,9).

Radiomics is a type of computer-assisted technology that can extract a large amount of features from medical images for automated analysis through high-throughput methods, enabling precise quantitative evaluation of lesions (10-12). These features provide deep insights into the microstructure, metabolism, and molecular expression of tumors, enabling more accurate evaluation of tumor malignancy and prognosis, as well as the objective quantification of tumor heterogeneity (13). The accuracy of radiomics models is influenced by lesion segmentation, which is a key factor (14,15). However, most

Highlight box

Key findings

- This study utilized data from 773 patients to develop and evaluate a combined model for predicting the nature of breast masses prior to Mammotome minimally invasive excisional biopsy based on semi-automatic segmentation. In the task of predicting tumor versus non-tumor patients, the final combined model achieved an area under the curve (AUC) of 0.890 (training cohort) and 0.780 (testing cohort). In the task of predicting adenosis versus other types of lesions, the final stacking model achieved an AUC of 0.890 (training cohort) and 0.771 (testing cohort).

What is known and what is new?

- Previous radiomics models for breast mass evaluation depend on manual lesion delineation, which is time-consuming and subjective.
- This study introduces two semi-automatic segmentation algorithms [DeepLabv3_ResNet50 and fully convolutional network (FCN)_ResNet50] that achieve high accuracy (Dice coefficients >92%), enabling robust radiomic feature extraction and improved predictive performance.

What is the implication, and what should change now?

- The constructed combined model can reduce reliance on manual delineation, enhance reproducibility, and effectively improve the diagnostic accuracy for the nature of breast masses.
- Further research and validation of the combined model in various clinical settings are necessary to support its integration into routine clinical practice and optimize patient-specific treatment plans.

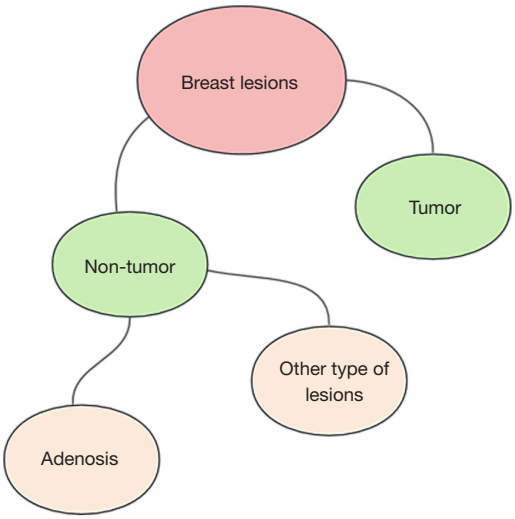

Figure 1 Binary classification tasks of the study.

current research still relies on professional radiologists to perform manual segmentation of each slice (16). Manual segmentation remains labor-intensive and prone to human subjectivity, often delaying diagnostic workflows and compromising feature reproducibility (17,18). Semi-automatic methods address these limitations by combining clinician oversight with algorithmic precision, ensuring both efficiency and consistency (19).

DeepLabv3_ResNet50 excels in capturing multi-scale contextual features through its atrous spatial pyramid pooling (ASPP) module, enabling precise boundary delineation of heterogeneous lesions (20,21). However, its computational complexity may limit real-time applications in resource-constrained settings. In contrast, fully convolutional network (FCN)_ResNet50 leverages a fully convolutional architecture to efficiently process entire US images, achieving higher global accuracy but occasionally overlooking subtle texture variations in low-contrast regions (22,23). Both segmentation models have unique strengths and advantages. We evaluate the accuracy of semi-automatic segmentation of US images using both models, as well as their effectiveness in downstream tasks for distinguishing the nature of breast lesions prior to minimally invasive resection. Ultimately, a combined model will be developed to further improve the performance of downstream models. We present this article in accordance with the TRIPOD reporting checklist (available at https://gs.amegroups.com/article/view/10.21037/gs-24-440/rc).

Methods

Study participants

Patients undergoing US-guided Mammotome-assisted minimally invasive resection at the Fifth Affiliated Hospital of Sun Yat-sen University from November 2018 to November 2023 were selected. The inclusion criteria were as follows: (I) completing minimally invasive breast surgery at our hospital; (II) preoperatively underwent US examination within 2 weeks; (III) complete clinical, imaging, and pathological data. The exclusion criteria were as follows: (I) insufficient or lack of US data for US images; (II) malignancy in other parts of the body. *Figure 1* shows the binary classification tasks of the study. A total of 773 patients were ultimately included (Figure S1) and divided into a training cohort and a test cohort. Among the 773 patients ultimately enrolled in the study, the non-tumor patients were also stratified into training and validation cohorts in a ratio of 8:2. And the clinical data were reviewed and collected from the electronic medical record system, and *Figure 2* displays the study design and pipeline. The sample size estimation for this study was performed using GPower 3.1 software. Based on the research background, we assumed a medium effect size for the relationship between radiomic features and clinical outcomes (Cohen's d=0.5), with a significance level of 0.05 and statistical power set to 0.8. According to these parameters, GPower estimated that a total sample size of 128 participants would be required. In this retrospective study, 773 patients were included, with 543 tumor patients and 230 non-tumor patients, providing sufficient statistical power for the analysis.

Ethical statement

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013), and has been approved by the Ethics Review Committee of The Fifth Affiliated Hospital, Sun Yat-sen University (reference number: K271-1). Since this is a retrospective analysis, the requirement for patient's informed consent was waived.

US acquisition

All patients enrolled underwent preoperative breast US examination using equipment including CHISON Q8 (Nantong, China), TOSHIBA aplio500 (Tokyo, Japan), GE

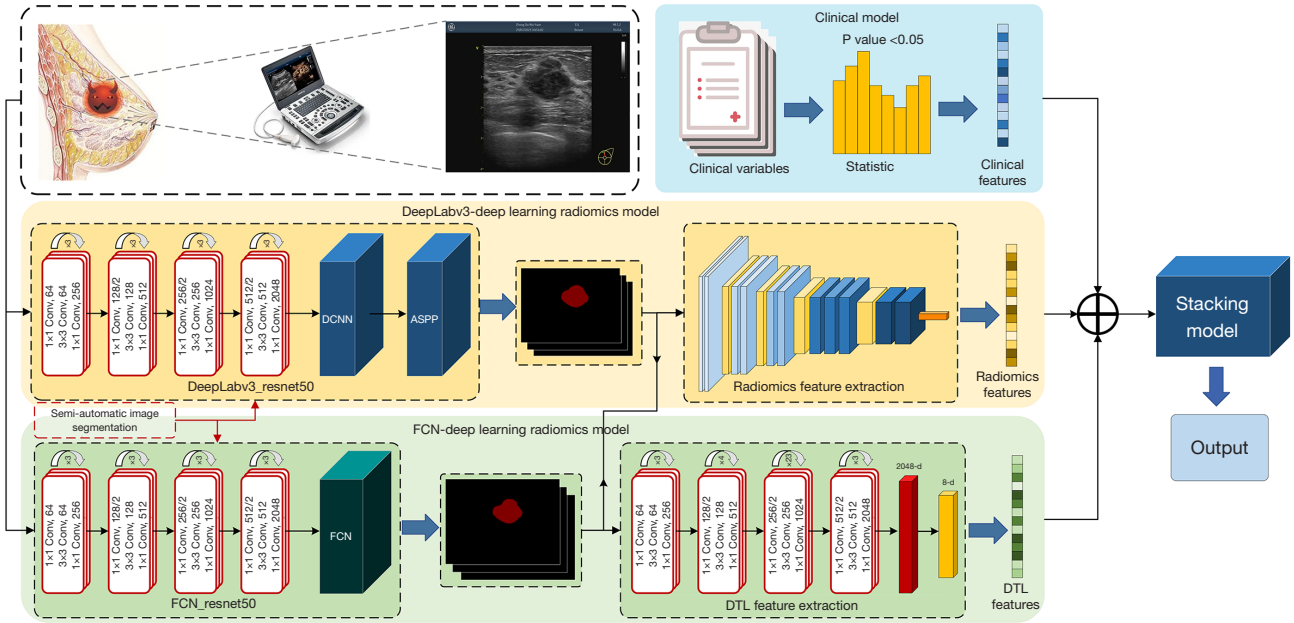

Figure 2 The study design and workflow of model development. ASPP, atrous spatial pyramid pooling; Conv, convolutional layer; DCNN, deep convolutional neural network; DTL, deep transfer learning; FCN, fully convolutional network.

LOGIQ E9 (Boston, USA), Philip EPIQ 7 (Amsterdam, Netherlands) and SIEMENS ACUSON Sequoia (Erlangen, Germany). Typically, patients adopt a supine position and fully expose the breast containing the lesion to capture and store the US images of the breast lesion.

Mammotome-assisted minimally invasive resection guided by US

All minimally invasive surgical procedures were performed using the "EnCor Enspire™ Breast Biopsy System (SenoRX Inc. E4230)". This system includes a 7G Mammotome biopsy device, control handle, vacuum suction pump, and related software. Patients were placed in a supine or semi-lateral position, with their arm raised. Moderate anesthesia (local anesthesia, 1% lidocaine ≤200 mg) was administered in the planned surgical area, which was the subcutaneous space behind the breast. A 3-mm incision was made in the designated location to insert the 7G Mammotome needle. Through US guidance, the needle was placed in an appropriate position on the deep surface of the breast lesion, ensuring that the lesion was fully within the needle's groove. Multiple rotational cuts were made to extract lesion tissue until no residual lesion was detected on US images. The surgical area was then subjected to hemostasis. All patients required compression bandages for 72 hours after surgery. *Figure 3* displays the surgical procedures.

Semi-automatic segmentation of region of interest (ROI)

Two radiologists, with over three years of experience and blinded to the pathology, utilized Picture Archiving and Communication Systems (PACS) to review and select US images of patients for inclusion in the study. We employed a semi-automated image segmentation method, utilizing the labelme assistive software and including DeepLabv3_ResNet50 and FCN_ResNet50 as segmentation models.

Feature extraction and screening

We imported the semi-automatic segmentation results of DeepLabv3_ResNet50 and FCN_ResNet50 into the PyRadiomics platform for radiomics feature extraction. After extracting features, we evaluated the repeatability and stability of radiomics parameters using intraclass correlation coefficients (ICC). We identified highly correlated radiomics features from the *t*-test with a P value of ≥0.05. Subsequently, we used the Spearman rank correlation test to evaluate the linear correlation between individual

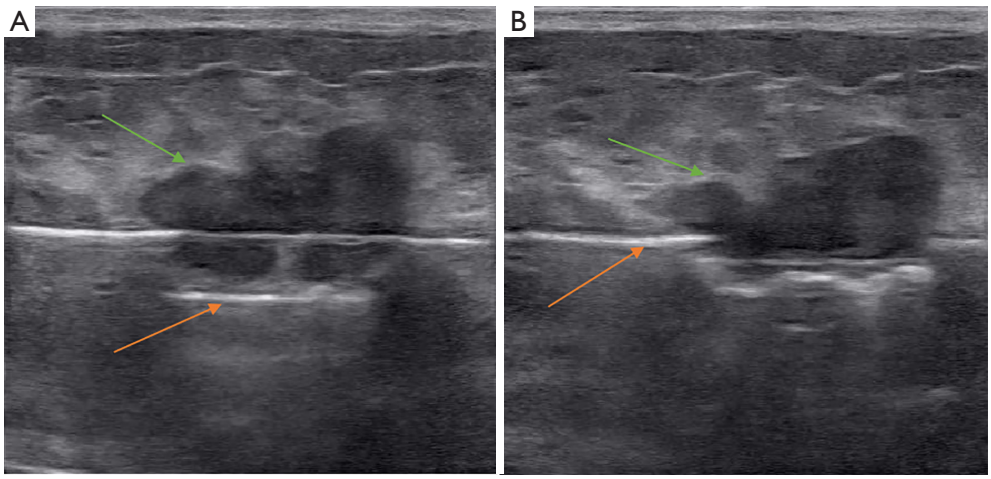

Figure 3 Mammotome-assisted minimally invasive resection guided by ultrasound. (A) Under ultrasound guidance, the needle was placed underneath the deep surface of the breast lump, and then the rotating cutter was opened to position the breast lump precisely in the needle's channel. (B) Closing the rotating cutter to remove the aspirated pathological tissue. The green arrow indicates a breast lesion, and the orange arrow indicates the rotating cutter.

features to eliminate redundancy. A previous study reported that features with strong correlation had higher absolute values of their correlation coefficients (24). Additionally, we selected one feature when the Spearman correlation coefficient between each feature was >0.9. Finally, in our previous work, we used least absolute shrinkage and selection operator (LASSO) regression for feature selection, where non-zero coefficients were considered valuable predictors within each feature group.

In this study, we selected ResNet101 as the pre-trained model and performed feature extraction on the dataset (25). The specific network architecture is described in *Figure 2*. After training the deep learning model, we extracted features from the avgpool layer as deep learning features. As the dimension of the transferred deep features is 2048, we used principal component analysis (PCA) to reduce their dimension and ensure balance between features. To improve the accuracy of the prediction model, we have integrated the radiomics features from the DeepLabv3_ResNet50 pathway and FCN_ResNet50 pathway, as well as 8-dimensional deep transfer learning (DTL) features, to construct a deep learning radiomics model.

Model construction and assessment

Using Python (version 3.12) for model building and evaluation, we selected features for clinical model building based on baseline statistics with P values less than 0.05. We screened and built radiomics features using different machine learning models into DeepLabv3-radiomics and FCN-radiomics models. The signature for building deep learning models were compressed DTL features. Our goal was to create deep learning radiomics labels, build DeepLabv3-deep learning radiomics model and FCN-deep learning radiomics model based on the selected radiomics features and compressed DTL features. Finally, we combined the DeepLabv3-deep learning radiomics model, FCN-deep learning radiomics model, and clinical model in a stacking manner to form the final model.

Statistical analysis

Using the SPSS software package (version 20.0) to evaluate baseline data, continuous variables were described as mean ± standard deviation, and categorical variables were described as frequency and percentage. For continuous variables, comparisons were made using independent sample *t*-test or Mann-Whitney *U* test; for categorical variables, comparisons were made using Chi-squared test or Fisher's exact test. The area under the curve (AUC) was compared using DeLong's test. A P value <0.05 was considered statistically significant (26). Optimal classification thresholds were determined by maximizing Youden's index (J = sensitivity + specificity – 1) to ensure clinical relevance.

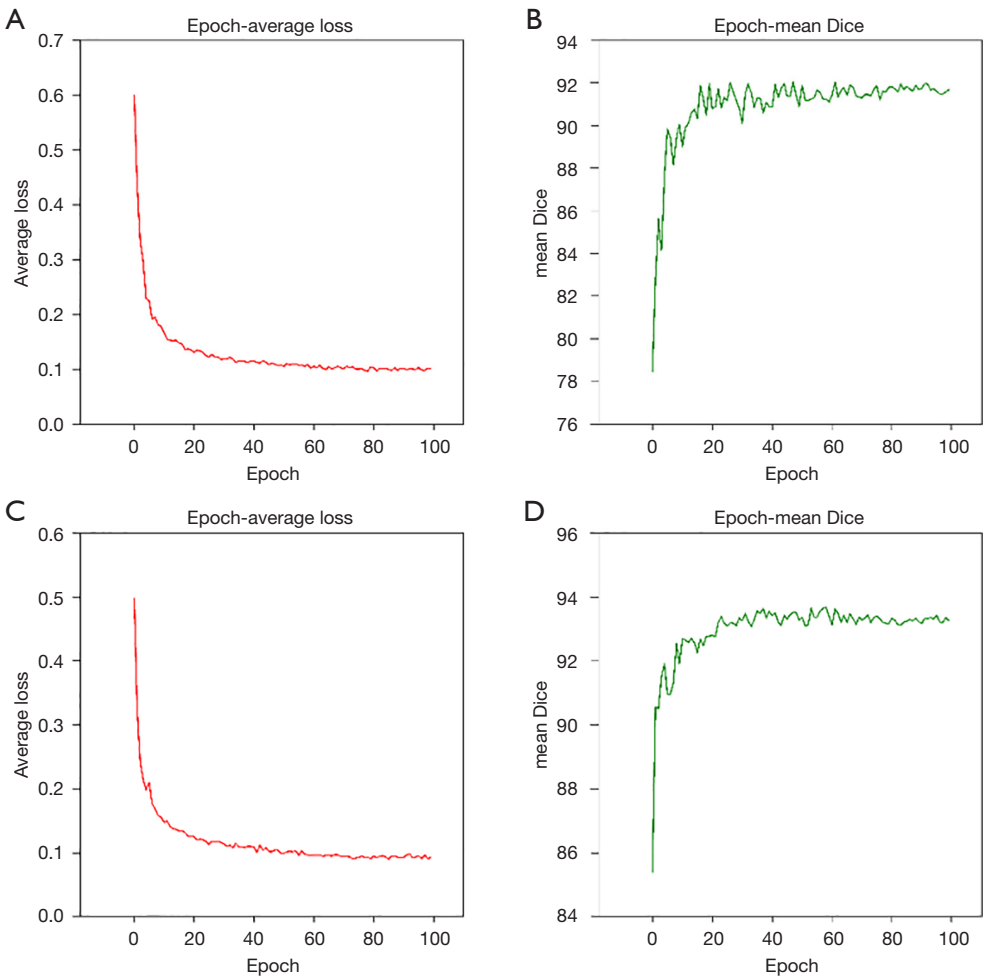

Figure 4 Training process of the segmentation models. (A) Epoch-average loss of DeepLabv3. (B) Epoch-mean Dice of DeepLabv3. (C) Epoch-average loss of FCN. (D) Epoch-mean Dice of FCN. FCN, fully convolutional network.

Results

Characteristics of patients

In this study, we recruited a total of 773 patients, including 543 with breast tumors and 230 with non-breast tumors. The characteristics of all patients are in the Table S1. Significant differences were observed in the clinical features between the two cohorts, including age, maximum diameter, and Breast Imaging Reporting and Data System (BI-RADS) classification. However, there were no significant differences in the tumor location and focus between the training and testing cohorts. After comprehensive multivariate analysis, age, maximum diameter, and BI-RADS classification were ultimately identified as key indicators and were rigorously integrated into the construction of our clinical prediction model. Table S2 presents a summary of characteristics for both the training and testing cohorts among non-breast tumor patients. Following comprehensive multivariate analysis, focus emerged as a crucial factor and was subsequently incorporated into the development of a clinical prediction model.

Results of semi-automatic segmentation

We summarized the training results of the two segmentation models (*Figure 4*). It was observed that both models achieved high accuracy. Specifically, as shown in *Table 1*, the DeepLabv3_ResNet50 model attained a global accuracy of 99.4%, an average intersection over union (IoU) of 86.2%, and an average Dice coefficient of 92.0%

Table 1 Best epoch of DeepLabv3_ResNet50 and FCN_ ResNet50

Model name	Global_acc, %	MIoU, %	Dice, %	MDice, %	Epoch
DeepLabv3	99.4	86.2	84.4–99.7	92.0	100
FCN	99.5	88.7	87.6–99.8	93.7	100

MIoU, median intersection over union; FCN, fully convolutional network.

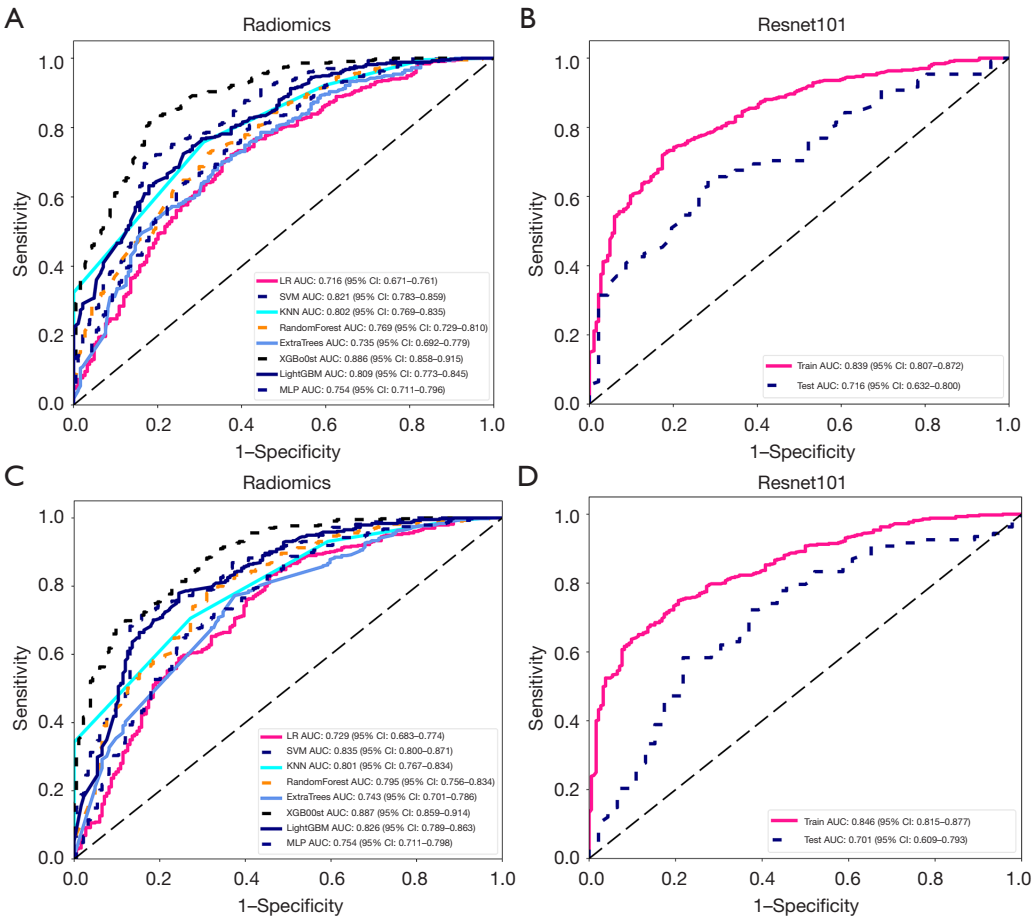

Figure 5 Radiomics models and deep learning models of DeepLabv3 and FCN. (A) ROC curve of DeepLabv3-radiomics model; (B) ROC curve of DeepLabv3-deep learning model; (C) ROC curve of FCN-radiomics model; (D) ROC curve of FCN-deep learning model. AUC, area under the ROC curve; CI, confidence interval; FCN, fully convolutional network; GBM, gradient boosting machine; KNN, K-nearest neighbors; LR, logistic regression; MLP, multi-layer perceptron; ROC, receiver operating characteristic; SVM, support vector machine.

during the best epoch. Similarly, the FCN_ResNet50 model attained a global accuracy of 99.5%, an average IoU of 88.7%, and an average Dice coefficient of 93.7% during the best epoch. Both segmentation models were trained for 100 epochs.

Performance of the radiomics and deep learning models in predicting tumor and non-tumor patients

In this study, we constructed radiomics models and deep learning models using the results of the DeepLabv3 and FCN approaches, respectively (*Figure 5*). The results

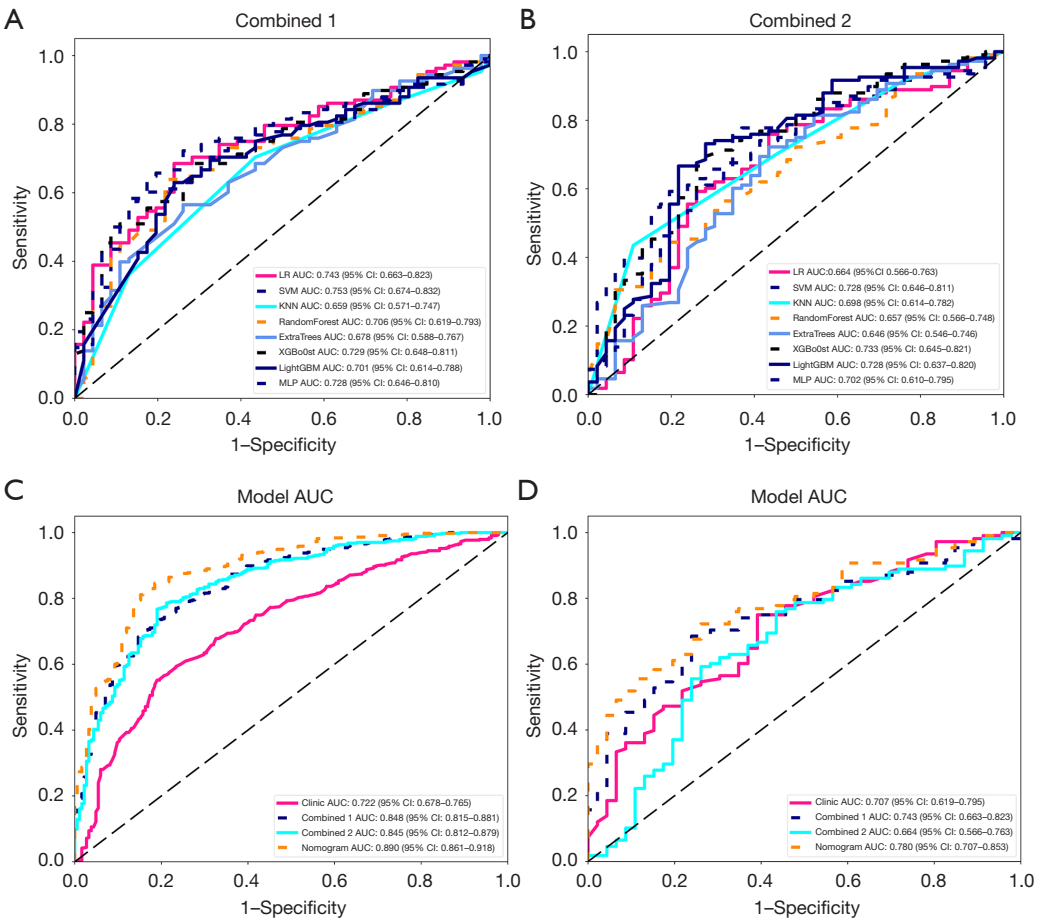

Figure 6 ROC curve of different models. (A) ROC curve of DeepLabv3 deep learning radiomics model; (B) ROC curve of FCN deep learning radiomics model; (C) ROC curve of stacking model in training cohort; (D) ROC curve of stacking model in testing cohort. Combined 1: DeepLabv3 deep learning radiomics model; combined 2: FCN deep learning radiomics model. AUC, area under the ROC curve; CI, confidence interval; FCN, fully convolutional network; GBM, gradient boosting machine; KNN, K-nearest neighbors; LR, logistic regression; MLP, multi-layer perceptron; ROC, receiver operating characteristic; SVM, support vector machine.

showed that the AUC range of the DeepLabv3-radiomics model was 0.716–0.886. The AUC of the DeepLabv3-deep learning model in the training cohorts was 0.839 [95% confidence interval (CI): 0.807–0.872], and in the testing cohorts, it was 0.716 (95% CI: 0.632–0.800). The AUC range of the FCN-radiomics model was 0.729–0.887. The AUC of the FCN-deep learning model in the training cohorts was 0.846 (95% CI: 0.815–0.877), and in the testing cohorts, it was 0.701 (95% CI: 0.609–0.793).

Development and performance of the combined and stacking models in predicting tumor and non-tumor patients

In the subsequent study, we combined deep learning features and radiomics features from different segmentation approaches to construct the DeepLabv3 deep learning radiomics model and FCN deep learning radiomics model. *Figure 6* shows the performance of these two models in the testing cohort, with an AUC of 0.659–0.753 for the

Table 2 Tumor *vs.* non-tumor classification

Model	Threshold	AUC (95% CI)	Sensitivity	Specificity
Stacking model (training)	0.42	0.890 (0.861–0.918)	0.844	0.815
Stacking model (testing)	0.38	0.780 (0.707–0.853)	0.713	0.739

AUC, area under the curve; CI, confidence interval.

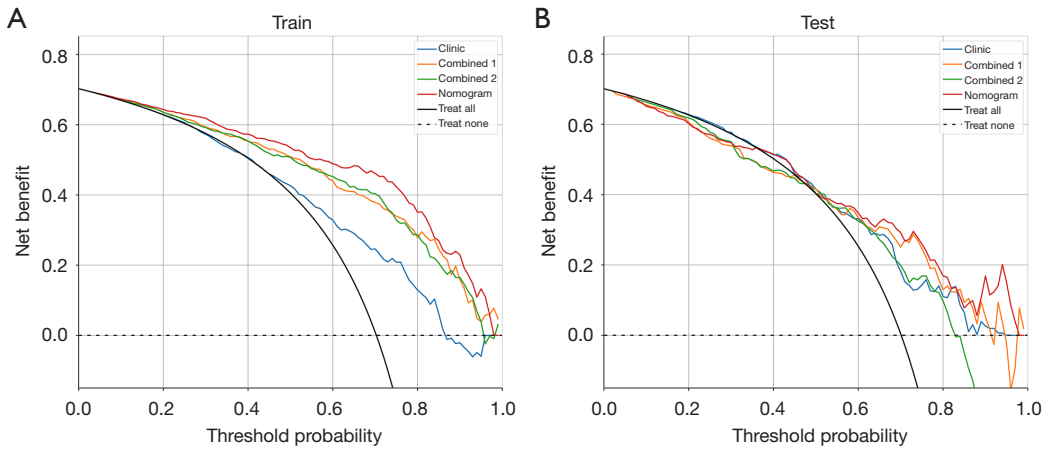

Figure 7 Decision curve analysis curves comparing tumor versus non-tumor prediction models. Plots show the decision curves of clinical model, DeepLabv3 deep learning radiomics model (combined 1), FCN deep learning radiomics model (combined 2), and the stacking model in the training (A) and testing cohorts (B). FCN, fully convolutional network.

DeepLabv3 model and an AUC of 0.646–0.733 for the FCN model. The DeepLabv3 model performed slightly better than the FCN model in the testing cohort.

Finally, we utilized stacking to fuse the clinical model, DeepLabv3 deep learning radiomics model, and FCN deep learning radiomics model using logistic regression, resulting in the final stacking model (nomogram). Experimental results in *Figure 6* demonstrated that the stacking model, which combined the clinical model, combined 1, and combined 2, significantly improved the ability to differentiate between tumor and non-tumor patients, with an AUC of 0.890 (95% CI: 0.861–0.918). This model demonstrated a sensitivity of 0.844 and a specificity of 0.815. In the testing cohort, the AUC of this stacking model reached 0.780 (95% CI: 0.707–0.853). The sensitivity of the stacking model was 0.713, and the specificity (0.739) was sufficiently high to identify non-tumor patients (*Table 2*).

To assess the clinical utility of different prediction models, we evaluated their performance using decision curve analysis (DCA). As shown in the *Figure 7*, various deep learning radiomics models, including the DeepLabv3 deep learning radiomics model (combined 1), the FCN deep learning radiomics model (combined 2), and the stacking model, were demonstrated in the DCA across the training and testing cohorts. These findings indicate that within the low threshold range (0.2–0.4), the stacking model effectively identifies most patients who require further evaluation or treatment, thereby reducing missed diagnoses. In the intermediate threshold range (0.4–0.6), the model demonstrates a favorable balance between sensitivity and specificity, making it well-suited for clinical decision-making for patients at moderate risk. Conversely, in the high threshold range (>0.6), the stacking model exhibits relatively weak performance for stringent tumor screening, suggesting that additional clinical information may be needed to support decision-making.

Development and performance of the combined and stacking models in predicting adenosis and other type of lesions

In the prediction task of adenosis and other lesion types, we have developed corresponding DeepLabv3 and FCN deep learning radiomics models. As depicted in *Figure 8*,

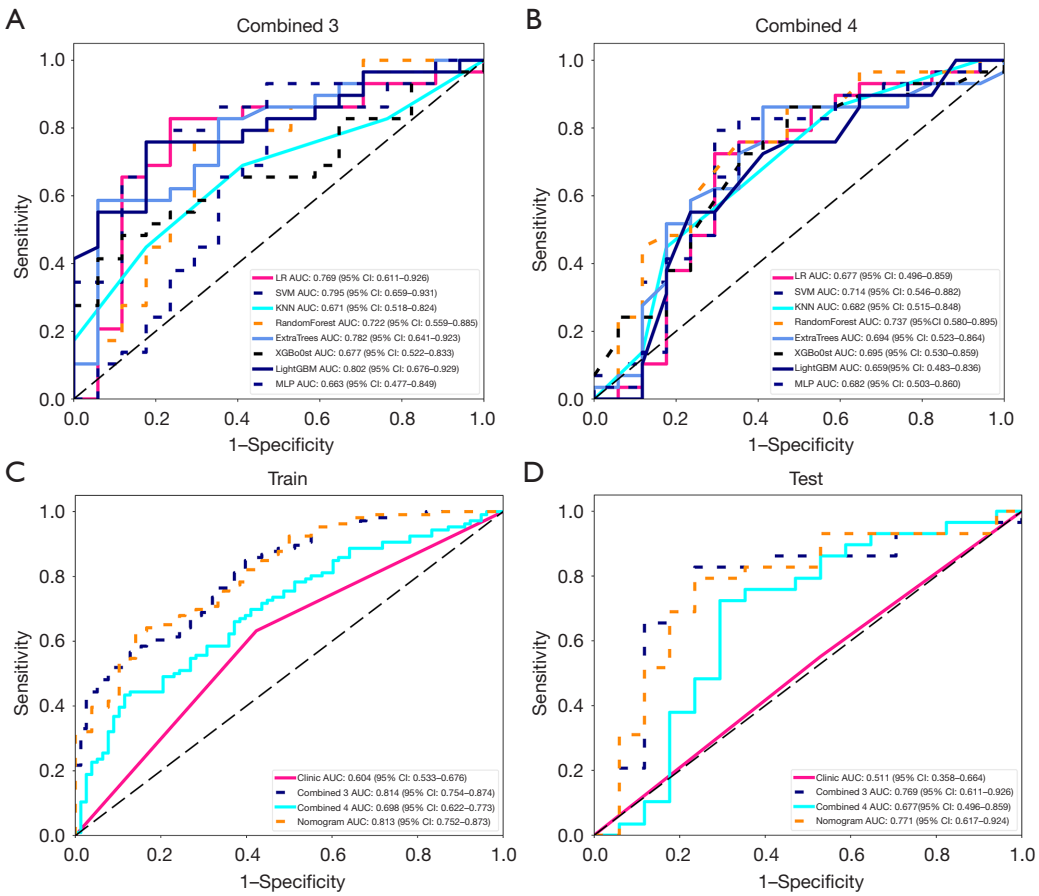

Figure 8 ROC curves of different models. (A) ROC curve of DeepLabv3 deep learning radiomics model; (B) ROC curve of FCN deep learning radiomics model; (C) ROC curve of stacking model in training cohort; (D) ROC curve of stacking model in testing cohort. Combined 3: DeepLabv3 deep learning radiomics model; combined 4: FCN deep learning radiomics model. AUC, area under the ROC curve; CI, confidence interval; FCN, fully convolutional network; GBM, gradient boosting machine; KNN, K-nearest neighbors; LR, logistic regression; MLP, multi-layer perceptron; ROC, receiver operating characteristic; SVM, support vector machine.

the performance of these two models in the testing cohort revealed AUC values ranging from 0.663 to 0.802 for the DeepLabv3 model and 0.677 to 0.737 for the FCN model. The final stacking model effectively distinguished patients with adenosis from those with other lesions, achieving an AUC of 0.813 (95% CI: 0.752–0.873) in the training cohort. The sensitivity and specificity of this model were 0.613 and 0.859, respectively. In the testing cohort, the stacking model exhibited an AUC of 0.771 (95% CI: 0.617–0.924), along with a sensitivity of 0.759 and specificity of 0.765 (*Table 3*).

The DCA curve presented in *Figure 9* demonstrated the clinical utility of the DeepLabv3 deep learning radiomics model, the FCN deep learning radiomics model, and the stacking model in both the training and testing cohorts. These findings suggest that in the low threshold range (0.1–0.3), the stacking model serves as an auxiliary tool for the preliminary screening of patients with adenopathy, and it should be used in conjunction with additional clinical information for a comprehensive evaluation. In the intermediate threshold range (0.3–0.6), when establishing a definitive diagnosis, the model effectively balances sensitivity and specificity, thereby reducing both missed and incorrect diagnoses. However, in the high threshold range (>0.6), where a high degree of diagnostic certainty is required, the model's performance may be insufficient, warranting cautious use.

Discussion

With the increase in breast disease screening coverage and

Table 3 Adenosis *vs.* other lesions classification

Model	Threshold	AUC (95% CI)	Sensitivity	Specificity
Stacking model (training)	0.35	0.813 (0.752–0.873)	0.613	0.859
Stacking model (testing)	0.40	0.771 (0.617–0.924)	0.759	0.765

AUC, area under the curve; CI, confidence interval.

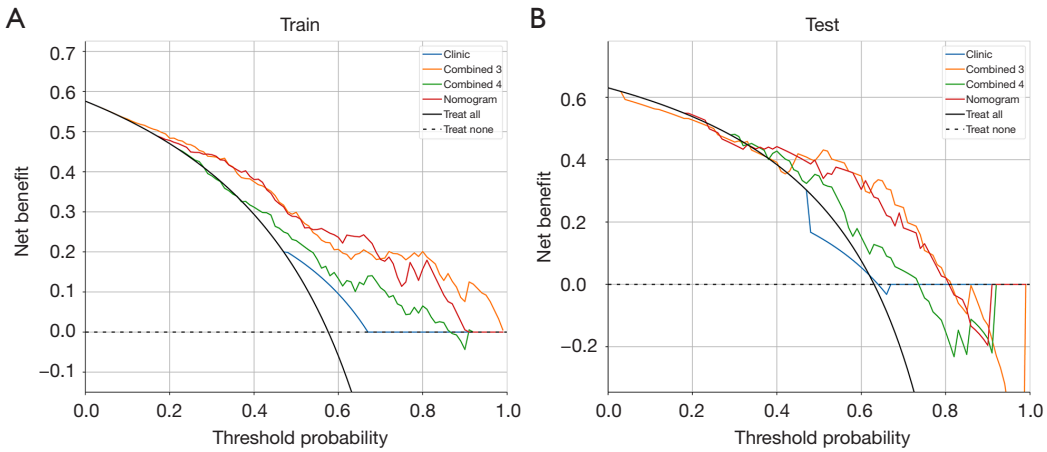

Figure 9 Decision curve analysis curves comparing adenosis versus other type of lesions prediction models. Plots show the decision curves of clinical model, DeepLabv3 deep learning radiomics model (combined 3), FCN deep learning radiomics model (combined 4), and the stacking model (nomogram) in the training (A) and testing cohorts (B). FCN, fully convolutional network.

public health awareness, the early detection rate of breast lesion has significantly improved (27). Given the presence of breast lesion, more and more patients face psychological pressure, such as cancerophobia, and therefore seek to determine the pathological nature of the lesion and undergo surgical removal (28). Providing appropriate treatment plans is crucial for surgeons. Research has shown that compared to traditional open surgery, Mammotome-assisted minimally invasive resection has significant advantages in terms of skin incision size, intraoperative blood loss, healing time, scar formation, wound infection, and cosmetic outcomes (29,30). Additionally, real-time dynamic monitoring of the lesion through US allows for more precise and complete removal. On the other hand, Mammotome-assisted minimally invasive resection is associated with minimal pain and high patient satisfaction (31). Preoperative assessment before surgery is helpful in achieving precision medicine, providing a basis for physicians to carry out individualized treatment, and significantly alleviating patients' psychological pressure.

Image segmentation is an important step in radiomics analysis and is crucial for identifying radiomic features. Common image segmentation methods include manual segmentation, semi-automatic segmentation, and automatic segmentation (32). Semi-automatic and automatic segmentation not only simplify the process but also improve repeatability. In recent years, deep learning has emerged as a powerful alternative for supervised segmentation (33-35). DeepLabv3_ResNet50 utilizes a self-attention encoder model structure and combines it with ResNet-50 as a feature extractor (36). It achieves high accuracy and efficiency while maintaining good scalability. FCN is a convolutional neural network used for image segmentation. It performs feature extraction and prediction on images layer by layer in a fully convolutional manner, ultimately obtaining pixel-level segmentation results. The FCN_ResNet50 algorithm also incorporates ResNet-50 as a feature extractor to utilize its excellent feature representation capability. Combined with the fully convolutional approach of FCN, it enables efficient image segmentation (37). The semi-automatic segmentation models used in this study demonstrate high accuracy, with a global accuracy of 99.4% (DeepLabv3_ResNet50) and 99.5% (FCN_ResNet50).

In the task of predicting tumor and non-tumor patients, the stacking model demonstrated impressive discriminatory

power in the training cohort, achieving an AUC value of 0.890 (95% CI: 0.861–0.918). When evaluated in the testing cohort, the model maintained a strong performance with an AUC of 0.780 (95% CI: 0.707–0.853). In the task of predicting adenosis and other lesion types, while the clinical model's predictive performance was suboptimal, the stacking model yielded superior results. Specifically, it achieved an AUC of 0.813 in the training cohort and 0.771 in the testing cohort. Although the clinical model underperformed in predicting adenosis (AUC =0.68), incorporating it into the ensemble improved specificity by 7%. This aligns with a study showing that even weak sub-models can enhance diagnostic robustness through complementary feature spaces (38). Compared with current clinical protocols, the stacking model based on semi-automatic segmentation not only effectively reduces unnecessary biopsies and follow-ups but also significantly decreases the manual annotation time required per case, thereby allowing clinicians to prioritize decision-making for repetitive tasks. Furthermore, the cost savings achieved with this technology enable healthcare institutions to reallocate resources to higher-priority care services, while the effective management of patients' preoperative anxiety underscores its psychosocial benefits.

Overall, our study demonstrates the efficacy of enhancing model performance through the integration of multiple predictive models. However, further work and research are necessary to confirm the reliability of these findings. The retrospective nature and single-institution data may constrain generalizability. While internal validation demonstrates model robustness under controlled conditions, future prospective, multicenter studies with heterogeneous cohorts are essential to assess real-world adaptability. We must continue to strive for improved research quality to ensure the accuracy and reliability of our findings.

Conclusions

In summary, this study utilized preoperative US images from Mammotome-assisted minimally invasive resections to develop an ensemble learning model employing a semi-automated segmentation approach. This model accurately distinguishes between tumor and non-tumor patients, as well as adenosis and other lesion types, thereby offering valuable insights for individualized treatment planning. The findings underscore the potential of this approach to enhance preoperative diagnostic accuracy and to guide personalized medical interventions. The stacking model demonstrates potential clinical value in accurately differentiating tumor from non-tumor patients and adenosis from other lesions. Future studies should focus on refining this model and expanding its applicability to additional types of breast lesions. Moreover, continuous improvements in research quality will be essential to ensure the clinical utility of these predictive models.

Acknowledgments

We thank the Department of Ultrasound for providing ultrasound images and appreciate the Python Technology providing the OnekeyAI platform, which greatly supports our research.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://gs.amegroups.com/article/view/10.21037/gs-24-440/rc

Data Sharing Statement: Available at https://gs.amegroups.com/article/view/10.21037/gs-24-440/dss

Peer Review File: Available at https://gs.amegroups.com/article/view/10.21037/gs-24-440/prf

Funding: None.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://gs.amegroups.com/article/view/10.21037/gs-24-440/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013), and has been approved by the Ethics Review Committee of The Fifth Affiliated Hospital, Sun Yat-sen University (reference number: K271-1). Since this is a retrospective analysis, the requirement for patient's informed consent was waived.

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;71:209-49.
2. Harbeck N, Gnant M. Breast cancer. Lancet 2017;389:1134-50.
3. Xiong X, Zheng LW, Ding Y, et al. Breast cancer: pathogenesis and treatments. Signal Transduct Target Ther 2025;10:49.
4. Ohuchi N, Suzuki A, Sobue T, et al. Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial. Lancet 2016;387:341-8.
5. Tadesse GF, Tegaw EM, Abdisa EK. Diagnostic performance of mammography and ultrasound in breast cancer: a systematic review and meta-analysis. J Ultrasound 2023;26:355-67.
6. Huang R, Lin Z, Dou H, et al. AW3M: An auto-weighting and recovery framework for breast cancer diagnosis using multi-modal ultrasound. Med Image Anal 2021;72:102137.
7. Banys-Paluchowski M, Rubio IT, Karadeniz Cakmak G, et al. Intraoperative Ultrasound-Guided Excision of Non-Palpable and Palpable Breast Cancer: Systematic Review and Meta-Analysis. Ultraschall Med 2022;43:367-79.
8. Wang H, Wang Q, Zhang Y, et al. Value of ultrasound BI-RADS classification in preoperative evaluation of the ultrasound-guided Mammotome-assisted minimally invasive resection of breast masses: A retrospective analysis. Exp Ther Med 2023;25:143.
9. Trayes KP, Cokenakes SEH. Breast Cancer Treatment. Am Fam Physician 2021;104:171-8.
10. Ab Mumin N, Ramli Hamid MT, Wong JHD, et al. Magnetic Resonance Imaging Phenotypes of Breast Cancer Molecular Subtypes: A Systematic Review. Acad Radiol 2022;29 Suppl 1:S89-S106.
11. Conti A, Duggento A, Indovina I, et al. Radiomics in breast cancer classification and prediction. Semin Cancer Biol 2021;72:238-50.
12. Sohn JH, Fields BKK. Radiomics and Deep Learning to Predict Pulmonary Nodule Metastasis at CT. Radiology 2024;311:e233356.
13. Tagliafico AS, Piana M, Schenone D, et al. Overview of radiomics in breast cancer diagnosis and prognostication. Breast 2020;49:74-80.
14. Yue WY, Zhang HT, Gao S, et al. Predicting Breast Cancer Subtypes Using Magnetic Resonance Imaging Based Radiomics With Automatic Segmentation. J Comput Assist Tomogr 2023;47:729-37.
15. Wang X, Xie T, Luo J, et al. Radiomics predicts the prognosis of patients with locally advanced breast cancer by reflecting the heterogeneity of tumor cells and the tumor microenvironment. Breast Cancer Res 2022;24:20.
16. Gryska E, Schneiderman J, Björkman-Burtscher I, et al. Automatic brain lesion segmentation on standard magnetic resonance images: a scoping review. BMJ Open 2021;11:e042660.
17. Fel JT, Ellis CT, Turk-Browne NB. Automated and manual segmentation of the hippocampus in human infants. Dev Cogn Neurosci 2023;60:101203.
18. Li Y, Liu Y, Huang L, et al. Deep weakly-supervised breast tumor segmentation in ultrasound images with explicit anatomical constraints. Med Image Anal 2022;76:102315.
19. Wang X, Bao N, Xin X, et al. Automatic evaluation of endometrial receptivity in three-dimensional transvaginal ultrasound images based on 3D U-Net segmentation. Quant Imaging Med Surg 2022;12:4095-108.
20. Murugappan M, Bourisly AK, Prakash NB, et al. Automated semantic lung segmentation in chest CT images using deep neural network. Neural Comput Appl 2023;35:15343-64.
21. Shia WC, Hsu FR, Dai ST, et al. Semantic Segmentation of the Malignant Breast Imaging Reporting and Data System Lexicon on Breast Ultrasound Images by Using DeepLab v3. Sensors (Basel) 2022;22:5352.
22. Gómez-Flores W, Coelho de Albuquerque Pereira W. A comparative study of pre-trained convolutional neural networks for semantic segmentation of breast tumors in ultrasound. Comput Biol Med 2020;126:104036.
23. Hsieh YH, Hsu FR, Dai ST, et al. Incorporating the Breast Imaging Reporting and Data System Lexicon with a Fully Convolutional Network for Malignancy Detection on Breast Ultrasound. Diagnostics (Basel) 2021;12:66.
24. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med 2018;18:91-3.
25. Toseef M, Olayemi Petinrin O, Wang F, et al. Deep transfer learning for clinical decision-making based on high-throughput data: comprehensive survey with

benchmark results. Brief Bioinform 2023;24:bbad254.
26. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.
27. Wang Y, Li Y, Song Y, et al. Comparison of ultrasound and mammography for early diagnosis of breast cancer among Chinese women with suspected breast lesions: A prospective trial. Thorac Cancer 2022;13:3145-51.
28. Dinapoli L, Colloca G, Di Capua B, et al. Psychological Aspects to Consider in Breast Cancer Diagnosis and Treatment. Curr Oncol Rep 2021;23:38.
29. Johnson AT, Henry-Tillman RS, Smith LF, et al. Percutaneous excisional breast biopsy. Am J Surg 2002;184:550-4; discussion 554.
30. Tang X. Mammotome-Assisted Liposuction: A Novel Technique for Accessory Breasts. Aesthetic Plast Surg 2017;41:517-23.
31. Chang DH, Shu YL. Clinic efficacy and safety of ultrasound-guided Mammotome-assisted surgery for patients with breast benign tumors. Eur Rev Med Pharmacol Sci 2023;27:5985-92.
32. Ilesanmi AE, Chaumrattanakul U, Makhanov SS. Methods for the segmentation and classification of breast ultrasound images: a review. J Ultrasound 2021;24:367-82.
33. Wang P, Chen P, Yuan Y, et al. Understanding Convolution for Semantic Segmentation. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA; 2018:1451-60.
34. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:2481-95.
35. Kwak D, Choi J, Lee S. Rethinking Breast Cancer Diagnosis through Deep Learning Based Image Recognition. Sensors (Basel) 2023;23:2307.
36. Zhao D, Che NY, Song ZG, et al. Pathological diagnosis of lung cancer based on deep transfer learning. Zhonghua Bing Li Xue Za Zhi 2020;49:1120-5.
37. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640-51.
38. Hägele M, Eschrich J, Ruff L, et al. Leveraging weak complementary labels enhances semantic segmentation of hepatocellular carcinoma and intrahepatic cholangiocarcinoma. Sci Rep 2024;14:24988.

Cite this article as: Huang Z, Zhu Q, Li Y, Wang K, Zhang Y, Zhong Q, Li Y, Zeng Q, Zhong H. Development and validation of a semi-automatic radiomics ensemble model for preoperative evaluation of breast masses in mammotome-assisted minimally invasive resection. Gland Surg 2025;14(3):391-404. doi: 10.21037/gs-24-440