scientific reports

Check for updates

OPEN

Predicting pathological complete response to breast cancer neoadjuvant therapy using multi-combination machine learning models based on vision transformer features

Linyong Wu[1,3], Songhua Li[1,2,3], Feng Chen[1,2,3], Chaojun Wu[1], Yan Lin[1], Shaofeng Wu[1], Dayou Wei[1]✉ & Xiaohong Xu[2]✉

The significance of multi-combination machine learning models utilizing vision transformer (VIT) features in forecasting pathological complete response (pCR) after breast cancer neoadjuvant therapy (NAT). A retrospective study was conducted on 124 breast cancer patients who were confirmed by biopsy pathology and underwent surgical resection after NAT to evaluate pCR, and they were divided into a training cohort (*n* = 87) and a validation cohort (*n* = 37). Deep learning features were extracted from pre-biopsy ultrasound images based on VIT, ResNet50, and VGG16 convolutional neural network algorithms. Based on the Wilcoxon test, 9, 7, and 7 high-value features were identified from VIT, Resnet50, and VGG16 features, respectively. Using 12 machine learning algorithms, 111, 87, and 82 models were developed utilizing VIT, Resnet50, and VGG16 features. The Stepglm [forward], NaiveBayes, and glmBoost + Ridge ensemble algorithms achieved the highest area under the curve (AUCs). In both the training and validation cohorts, the AUCs of the optimal algorithms were: 0.872 and 0.839 for VIT, 0.837 and 0.804 for Resnet50, and 0.835 and 0.804 for VGG16. The predictive models developed based on VIT features have higher values in evaluating NAT-pCR for breast cancer compared to features quantified by other deep learning algorithms. The application of multi-combination models allows for the selection of the optimal algorithm to achieve higher prediction performance.

Keywords Breast cancer, Vision transformer, Deep learning, Machine learning, Neoadjuvant therapy

Breast cancer has become a cancer disease that imposes a global public burden. However, breast cancer is often found to have progressed to the middle or advanced stages, thereby losing the opportunity for surgical intervention. Neoadjuvant therapy (NAT) has become a standard conversion treatment strategy for middle and advanced breast cancer, aiming to reduce tumor staging, shrink tumor volume, alleviate lymph node metastasis, and create surgical opportunities for patients with middle and advanced stages[1]. Pathologic complete response (pCR) is the desired goal of NAT. Patients who achieve pCR can adopt breast-conserving strategies or even avoid surgical intervention. pCR is based on the pathological tissue of the tumor after surgery, which is evaluated by immunohistochemical techniques, including the Miller-Payne (MP) grading system and the residual cancer burden (RCB) grading system, to visually reflect the true changes in the lesions compared with the puncture tissue pathology[2,3]. However, some studies had shown that about 37% of breast cancer patients could not achieve the desired goal through NAT, and about 5% of them would therefore increase treatment costs or suffer adverse reactions[4]. Accurately identifying breast cancer patient populations that can benefit from NAT is currently a hot clinical challenge to avoid excessive NAT.

[1]Department of Medical Ultrasound, Maoming People's Hospital, Maoming 525000, Guangdong Province, P. R. China. [2]Department of Medical Ultrasound, The Affiliated Hospital of Guangdong Medical University, 524000 Zhanjiang, Guangdong Province, P. R. China. [3]Linyong Wu, Songhua Li and Feng Chen contributed equally. ✉email: weidayoumm@163.com; 13828297586@139.com

The imaging biomarker features of medical images have been proven to be correlated with NAT tumor response. For instance, changes in background parenchymal enhancement observed in breast MRI before and after NAT revealed that the pCR group exhibited significantly greater changes compared to the non-pCR group[5]; the ultrasound models developed based on posterior acoustic shadowing, margin, and calcification features of breast lesions predicted pCR, achieving an area under the curve (AUC) of 0.769, with sensitivity (SEN) and specificity (SPE) of 0.83 and 0.57, respectively[6]. Another model utilizing ultrasound size, posterior acoustic shadowing, and shape features predicted pCR, yielding an AUC of 0.758, SEN of 0.67, and SPE of 0.78[7]. Furthermore, the diagnostic performance of predicting NAT-pCR using the rising time parameter obtained from contrast-enhanced ultrasound imaging was good, with an AUC reaching 0.71[8]. Nevertheless, the predictive accuracy of such methods only attains moderate effectiveness, prompting the need to explore imaging biomarkers with higher utilization value. Deep learning (DL), an artificial intelligence (AI) technology capable of extracting imaging features representative of tumor heterogeneity from medical images, holds potential for achieving precision diagnosis and treatment[9]. Convolutional neural network (CNN) DL algorithms include ResNet50 and VGG16[10,11]. For example, the ResNet50 model could identify breast cancer patients who were likely to achieve pCR from NAT among 603 multi-center breast cancer ultrasound images, achieving an AUC of 0.88[12]. The pCR prediction models developed using VGG16 features from pathological images based on support vector machine demonstrated optimal performance, with an AUC of 0.79[13]. Recently, vision transformer (VIT) has become a more valuable DL algorithm for image utilization, which has been widely applied in image feature extraction and prediction tasks. For example, the predictive model developed using VIT algorithm based on CT images had been utilized to assess postoperative recurrence of lung cancer, achieving an AUC of 0.90[14]. Compared to CNN algorithms, VIT employs an attention mechanism to increase the weight of important image parts, potentially enabling better prediction of breast NAT-pCR. For instance, when VIT and seven CNN algorithms were used to develop pCR prediction models based on CT images, VIT features performed best in both training and validation cohorts, while VGG16 features performed the worst[15]. Although VIT has begun to be applied in breast cancer NAT-pCR prediction, studies based on ultrasound images remains scarce. Moreover, the current comparisons of various deep learning features often rely on a single machine learning algorithm to develop models, raising questions about whether the chosen algorithm is optimal and whether this may lead to biased comparison results. Cross-validation methods for feature selection and model development using various machine learning algorithms have been widely applied in exploring the applicability of models in the past, for instance, 12 machine learning algorithms demonstrated good performance in developing multi-combination models for predicting lymph node metastasis in cholangiocarcinoma based on radiomics features.

Based on the above background, the aim of this study is to explore the value of 113 multi-combination models generated by 12 machine learning algorithms, using ultrasound DL features extracted by VIT, ResNet50, and VGG16 algorithms, in predicting cancer NAT-pCR response. In addition, the clinical significance of VIT features was interpreted based on radiomics features.

Materials and methods

Ethical approval

The Ethics Committee of Maoming People's Hospital waived the need for obtaining informed consent because the study was retrospective. This retrospective study had obtained ethical approval from the Ethics Review Committee of Maoming People's Hospital (PJ2024MI-K040-01). All analyses were conducted in accordance with relevant guidelines and regulations, including the Helsinki Declaration.

Breast cancer patients undergoing NAT

This study initially recruited 501 patients with breast space-occupying lesions who underwent ultrasound-guided puncture histopathology biopsy at Maoming People's Hospital from January 2020 to December 2023. Inclusion Criteria: (1) Patients with primary breast space-occupying lesions classified as the Breast Imaging Reporting and Data System (BI-RADS) category 4 or 5 based on pre-biopsy ultrasonography. (2) Patients confirmed as breast cancer by puncture histopathology. (3) Patients who received NAT to obtain surgical opportunities, including chemotherapy, endocrine therapy, and targeted therapy. (4) Patients who underwent surgical resection after NAT to assess the tumor response. Exclusion Criteria: (1) Patients with bilateral lesions or multiple unilateral lesions. (2) Lesions that were too large to be fully captured or with suboptimal image quality for analysis. (3) Tumor response assessment not performed using the Miller & Payne (MP) grading system. (4) Patients with missing follow-up data or severing lack of clinical information (Fig. 1).

The clinical pathological data were sourced from the medical record system, encompassing age, menopausal status, lesion size, BI-RADS classification, white blood cell (WBC), hemoglobin count, fibrinogen, platelet distribution width (PDW), neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), systemic immune inflammation index (SII), pan-immune inflammation value (PIV), human epidermal growth factor receptor 2 (HER2), Ki67 expression[41]. A total of 124 patients with breast disease were finally included, aged from 32 to 82 years old, with an average age of (52.55 ± 9.41) years old. Among them, there were 23 cases of BI-RADS 4a/b, 51 cases of 4c, and 50 cases of 5. The overall cohort was randomly split into a training cohort ($n = 87$) and a validation cohort ($n = 37$) at a ratio of 7:3.

Tumor response assessment

Tumor response to NAT in breast cancer, assessed by the MP system, was graded as follows: Grade 1 (no change), Grade 2 (≤30% reduction), Grade 3 (30-90% reduction), Grade 4 (>90% reduction with scattered cells), and Grade 5 (no residual invasive cancer, with intraductal carcinoma remnants)[42]. Grades 1–3 were classified as non-pCR, while Grades 4–5 were considered as pCR.

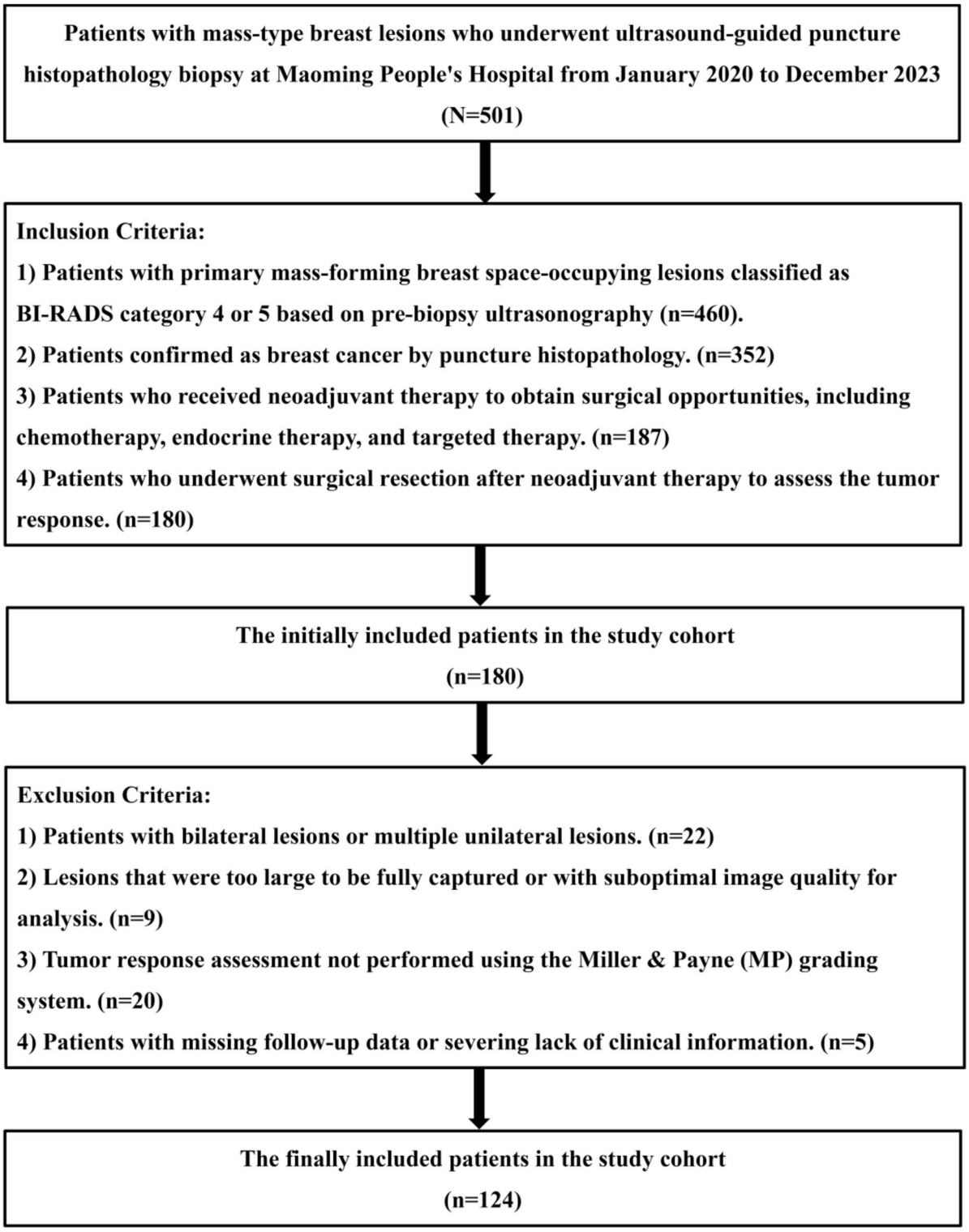

Fig. 1. Flow chart of study cohort.

Ultrasound image acquisition and deep learning feature analysis

The ultrasound physicians involved in image acquisition all possessed 3 years of experience in diagnosing breast lesions, ensuring that the captured images were the largest and clearest representations of the lesions. The BI-RADS classification was independently assigned by an ultrasound physician with 5 years of experience in diagnosing breast lesions, ensuring the accuracy and reliability of the diagnostic outcomes. The image acquisition process was facilitated by high-end ultrasound diagnostic systems, including Mindray Resona7, GE LOGIQ E9/10, and Philips EPIC7C, equipped with high-frequency linear array transducers (L11-3U, ML9L, L12-5, L12-

3) operating at frequencies ranging from 5 to 13 MHz. Appropriate machine parameters, such as gain, depth, acoustic window, mechanical index, and focal zone, were meticulously adjusted to optimize image quality. The acquired images of breast space-occupying lesions were saved in the Digital Imaging and Communications in Medicine (DICOM) format and subsequently loaded into the ITK-SNAP (3.80 version) software for further processing. An augmentation strategy involving random horizontal and vertical flipping was applied to the original ultrasound images to enhance their diversity. The regions of interest (ROIs) for breast space-occupying lesions were manually segmented in collaboration by two ultrasound physicians, each with 5 years of experience in breast diagnosis, using the ITK-SNAP software. The ComBat method was applied to calibrate for the variabilities in radiomic feature changes caused by differences in ultrasound equipment[43]. This was achieved by utilizing the PyRadiomics package to extract 1316 radiomics features from both the original ultrasound images and their corresponding transformed ROIs images, including: shape features, first-order statistical features, gray-level co-occurrence matrix (GLCM) features, gray-level run length matrix (GLRLM) features, gray-level size zone matrix (GLSZM) features, gray-level dependence matrix (GLDM) features, and neighboring gray tone difference matrix (NGTDM)[44]. With the ROI as the centerpiece, the smallest bounding rectangle encompassing the ROI's information was extracted and resized to 224 × 224 pixels, format was "PNG", serving as the input for the deep learning model. Supervised DL models, pre-trained on the ImageNet dataset, had been frequently leveraged for medical image analysis[45]. Consequently, from each ROIs image, deep learning features were extracted utilizing various deep learning models, including VIT, ResNet50, and VGG16, all of which were initialized with ImageNet pre-trained model parameters[46]. Given that deep learning harnesses image information at a scale of tens of thousands of dimensions, a compression process was applied to the image information, ultimately yielding 108 deep learning features[47]. These features encapsulate the essential information from the ROIs images, facilitating their utilization in downstream medical image analysis tasks (Fig. 2).

Development and validation of multi-combination machine learning models

The Z-score method initially standardized the overall cohort. The Wilcoxon test was then employed to identify high-value features that were indicative of favorable NAT-pCR. Subsequently, 12 distinct algorithms were leveraged to develop machine learning prediction models based on these high-value features, encompassing least absolute shrinkage and selection operator (LASSO), Ridge, elastic net (Enet), Stepglm, support vector machine (SVM), glmBoost, linear discriminant analysis (LDA), plsRglm, random forest (RF), gradient boosting machine (GBM), XGBoost, and naive bayes (NB)[37]. Within a cross-validation framework, the strategy was adopted wherein one algorithm was used for feature selection while another 11 algorithms were deployed to construct the classification prediction model, resulting in a maximum of 113 algorithm combinations. The best algorithms developed based on deep learning features were used to construct the radiomics machine learning models. The interpretation of clinical significance of deep learning features was based on correlated radiomics features. The receiver operating characteristic curve (ROC) was applied to evaluate the performance of the models, and AUC, SEN, and SPE were calculated. Calibration curve analysis and Hosmer-Lemeshow (HL) test were used to evaluate models calibration and goodness of fit. Decision Curve Analysis (DCA) was used to evaluate the net clinical benefit or utility of models.

Literature review on artificial intelligence based on ultrasound or pathology imaging

Before November 10, 2024, the literature was retrieved based on the PubMed database with the keyword "breast and neoadjuvant and (radiomics or deep learning)". The literature on tumor response assessment of NAT for breast cancer was comprehensively reviewed and sorted out in terms of deep learning study types. The main

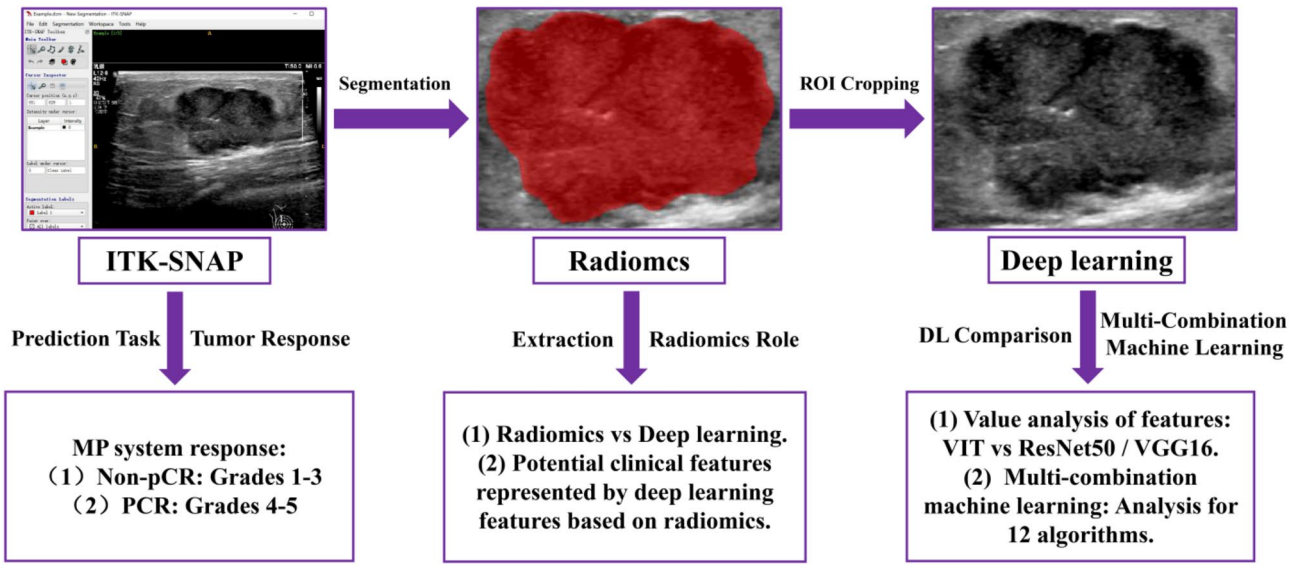

Fig. 2. Image feature extraction and analysis.

author, publication year, study population, tumor response standards, model type, AUC, SEN, SPE were extracted according to the full text of the literature. In addition, the literature based on pathological image analysis was also summarized and analyzed. A meta-analysis was conducted based on the Stata (version 12.0) software.

Statistical analysis

SPSS (17.0 version) and R (3.68 version) software were used for statistical analysis and plotting. Continuous variables were represented by mean values ± standard deviations, and t-test was used for comparison; categorical variables were represented by examples (%), and Chi square test was used for comparison. $P < 0.05$ was considered statistically significant.

Results

The baseline characteristics of breast cancer cohorts

44 cases (50.57%) in the training cohort showed pCR, while 19 cases (51.35%) in the validation cohort demonstrated non-pCR. The baseline characteristics of the two cohorts were summarized in Table 1. No significant differences in baseline characteristics were observed between the two cohorts.

Development and validation of multi-combination machine learning models

Among the overall cohort, 9, 7, and 7 high-value features were identified from 108 features extracted by the VIT, Resnet50, and VGG16 deep learning algorithms, respectively, based on the Wilcoxon test (Fig. 3). These features were then input into 113 combinations of algorithms, and predictive models consisting of 3 or more than 3 features were selected, resulting in 111, 87, and 82 models constructed from VIT, Resnet50, and VGG16 features, respectively (Fig. 4). The AUC of the validation cohort was used as the criterion for selecting the optimal model. For VIT, Resnet50, and VGG16 features, the highest AUCs were achieved in the Stepglm [forward], NB, and glmBoost + Ridge combination algorithms, respectively. The AUCs of the optimal algorithms in the training and validation cohorts were 0.872 and 0.839 for VIT, 0.837 and 0.804 for Resnet50, and 0.835 and 0.804 for VGG16. The HL test demonstrated a good consistency between the best algorithm models and pathological outcomes (VIT_Stepglm [forward], training cohort, $P = 0.45$, validation cohort, $P = 0.61$; Resnet50_NaiveBayes, training cohort, $P = 0.44$, validation cohort, $P = 0.88$; VGG16_glmBoost + Ridge, training cohort, $P = 0.50$, validation cohort, $P = 0.27$). DCA demonstrated that the VIT_Stepglm [forward] models possessed good potential clinical application value (Fig. 5). Among the three deep learning features, VIT features exhibited the best predictive performance, while Resnet50 and VGG16 features showed similar predictive effects. The Stepglm [forward] machine learning algorithm based on VIT features performed the best (AUC = 0.839), followed by VGG16 features (AUC = 0.787), and Resnet50 features performed the worst (AUC = 0.754). For the NB algorithm, Resnet50 features resulted in the best performance (AUC = 0.804), followed by VGG16 features (AUC = 0.678), and VIT features performed the worst (AUC = 0.667). Ridge algorithm achieved the best performance with VGG16 features (AUC = 0.804), followed by VIT features (AUC = 0.778), and Resnet50 features yielded the third-highest AUC of 0.772. Both the Stepglm (direction = forward) and Ridge algorithms demonstrated moderate predictive performance when applied to the quantitative features extracted from the three deep learning algorithms. Various other machine learning models also showed moderate performance in predicting favorable tumor responses to NAT for breast cancer, indicating the predictive stability of these models. Table 2 summarized the optimal combination algorithms for different machine learning algorithms.

Interpreting potential clinical significance represented by deep learning features based on radiomics

Twenty-six high-value features were identified from 1316 radiomics features based on the Wilcoxon test. Predictive models were developed using the optimal algorithms of lasso + Stepglm [forward], NB, and Ridge. The AUC, SEN, and SPE of the training cohort and validation cohort were 0.763, 0.80, 0.67 and 0.626, 1.00, 0.39; 0.694, 0.41, 0.95 and 0.722, 0.95, 0.56; 0.697, 0.75, 0.60 and 0.722, 0.89, 0.56, respectively (Fig. 6**A and 6B**). The correlation between the nine VIT deep learning features and the 26 radiomics features was presented in **Fig. 6C**. The results showed that 7 radiomics features were correlated with deep learning features, with correlation coefficients ranging from − 0.19 to 0.21. Among them, there were three GCLM features, two GLSZM features, one GLRLM feature, and one GLDM feature, representing the heterogeneity and complexity of texture distribution within ultrasound images in breast cancer.

Literature review on artificial intelligence based on ultrasound or pathology imaging

Among the 276 literature retrieved, 14 were deep learning literature and 7 were radiomics literature based on ultrasound imaging related to NA therapy for breast cancer, 7[3,4,12,17–20] and 5[17,21–23,27] of which were enrolled meta analysis, respectively. 8 literature on deep learning based on pathology imaging, of which 5[13,24–27] were included in the meta analysis (**Table 3**). The meta analysis revealed that the pooled AUC, SEN, and SPE of the deep learning ultrasound models were 0.91 (0.89–0.93), 0.86 (0.83–0.89), and 0.84 (0.78–0.88), respectively. In contrast, the pooled AUC, SEN, and SPE of the radiomics ultrasound models stood at 0.81 (0.78–0.85), 0.78 (0.68–0.85), and 0.72 (0.62–0.80), respectively. Interestingly, the deep learning pathological models achieved the pooled AUC of 0.81 (0.77–0.84), a SEN of 0.68 (0.59–0.77), and a SPE of 0.85 (0.72–0.92), indicating that the ultrasound models seemed to perform better in predicting breast cancer NAT tumor response compared to the pathological model (Fig. 7).

Characteristics	Training cohort ($n=87$)	Validation cohort ($n=37$)	*P*
Neoadjuvant therapy			
chemotherapy	42	16	0.09
endocrine	14	3	
chemotherapy + endocrine	0	2	
chemotherapy + targeted	31	16	
Age (year)			0.76
< 60	66	29	
≥ 60	21	8	
Menopause			0.05
no	20	3	
yes	67	34	
Tumor size (mm)			0.16
< 30	42	23	
≥ 30	45	14	
BI-RADS			0.26
4a/b	13	10	
4c	39	13	
5	35	14	
WBC (10^9/L)			0.81
< 6.13	42	17	
≥ 6.13	45	20	
PDW			0.07
< 14.5	64	21	
≥ 14.5	23	16	
Hemoglobin (g/L)			0.97
< 135.5	52	22	
≥ 135.5	35	15	
Fibrinogen (g/L)			0.25
< 2.92	35	19	
≥ 2.92	52	18	
NLR			0.07
< 1.55	22	4	
≥ 1.55	65	33	
PLR			0.24
< 130.66	28	16	
≥ 130.66	59	21	
SII			0.98
< 755.1	59	25	
≥ 755.1	28	12	
PIV			0.42
< 285	57	27	
≥ 285	30	10	
HER2			0.29
negative	49	17	
positive	38	20	
Ki67			0.12
< 30%	17	12	
≥ 30%	70	25	
Tumor response			0.94
Non-Responder	43	18	
Responder	44	19	

Table 1. The baseline characteristics of breast cancer cohorts.

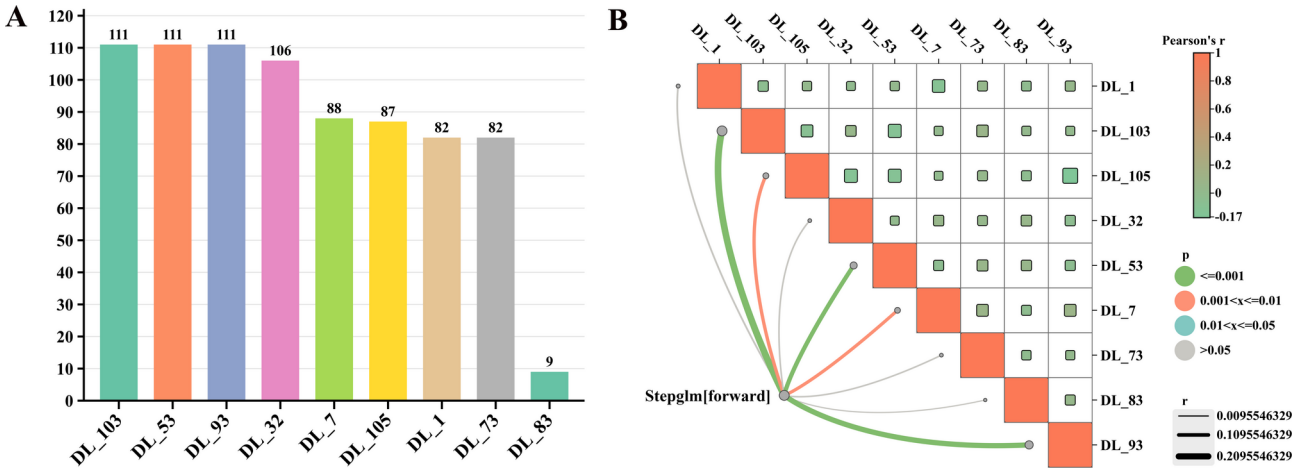

Fig. 3. Identification of high-value VIT features related to pCR. (**A**) Feature utilization rates across 111 machine learning models based on VIT features. (**B**) Heatmap of correlations among 9 VIT high-value features.

Discussion

Accurately identifying the population of breast cancer patients who benefit from NAT is conducive to minimizing unnecessary over-treatment. DL features have demonstrated high value in predicting NAT-pCR responses. This study aimed to explore the value of 113 multi-combination models generated by 12 machine learning algorithms, using ultrasound features extracted by VIT, ResNet50, and VGG16, in predicting breast cancer NAT-pCR responses. The results indicated that the predictive model developed based on VIT features had a higher value in evaluating the efficacy of NAT for breast cancer compared to features quantified by other deep learning methods. The application of multi-combination models allowed for the selection of the optimal algorithm to achieve higher prediction performance.

Literature review on deep learning in predicting NAT-pCR for breast cancer

AI technology serves as a groundbreaking force in transcending traditional image vision, enabling high-throughput quantification and enhanced utilization of heterogeneous imaging features in medical images, encompassing MRI, CT, ultrasound, and pathological images. High-throughput imaging features have been widely applied in predicting NAT-pCR in breast cancer. MRI-derived radiomics features had been extensively validated by numerous meta-analyses to exhibit favorable performance in NAT-pCR prediction, with pooled AUC values ranging from 0.78 to 0.85[28–31]. Meanwhile, MRI-derived DL features had shown even more promising results, achieving a pooled AUC of 0.92 through meta-analyses. Similarly, ultrasound-derived radiomics features, when subjected to a meta-analysis encompassing eight studies, exhibited a pooled AUC, SEN, and SPE of 0.86, 0.87, and 0.78, respectively[32]. However, the application value of ultrasound-based DL features in predicting NAT-pCR remained under-explored. The PubMed search uncovered 14 studies pertaining to the use of DL features in predicting NAT outcomes, with eight of these studies collectively demonstrating a pooled AUC of 0.91, SEN of 0.87, and SPE of 0.83, mirroring the outcomes observed with MRI-DL features and outperforming both MRI and ultrasound radiomics features, surpassing the meta-analysis of the four ultrasound studies included in this study. Furthermore, despite the potential of pathological image-based DL features, a comprehensive analysis of their utility has yet to be conducted. The search identified eight studies on pathological image-related DL features, of which five reported a pooled AUC, SEN, and SPE of 0.80, 0.68, and 0.85, respectively, suggesting room for improvement. This subpar performance could be attributed to the inherent complexity of pathological images and the maturity level of AI technologies employed. Notably, VIT features had already been explored for NAT-pCR prediction in CT and pathological images[15,24], underscoring their potential. Nevertheless, the application of VIT features to ultrasound images for this purpose remained an area ripe for further investigation.

Multi-combination machine learning models for predicting NAT-pCR

The utilization of DL feature extraction followed by their presentation to machine learning algorithms for developing predictive models has been widely embraced in medical imaging. Nevertheless, the question of which machine learning algorithm best suits DL features remains an area requiring further investigation. For instance, when developing predictive models using DL features extracted from chest CT images via algorithms such as ResNet50, ResNet101, AlexNet, VGG16, VGG19, GoogLeNet, SqueezeNet, Xception, and presenting these features to machine learning algorithms including SVM, k-nearest neighbors, RF, decision tree, and NB, it was observed that the combination of ResNet50 features and SVM machine learning algorithm yielded the best results, achieving an accuracy of 0.963[33]. Similarly, in another study where DL features were extracted from chest CT images using ResNet18 and GoogleNet, and presented to RF, SVM, fast decision tree, and NB algorithms for predictive model development, it was found that the ResNet18 features combined with SVM also demonstrated optimal performance[34]. This suggested that SVM algorithms may be more compatible with ResNet neural

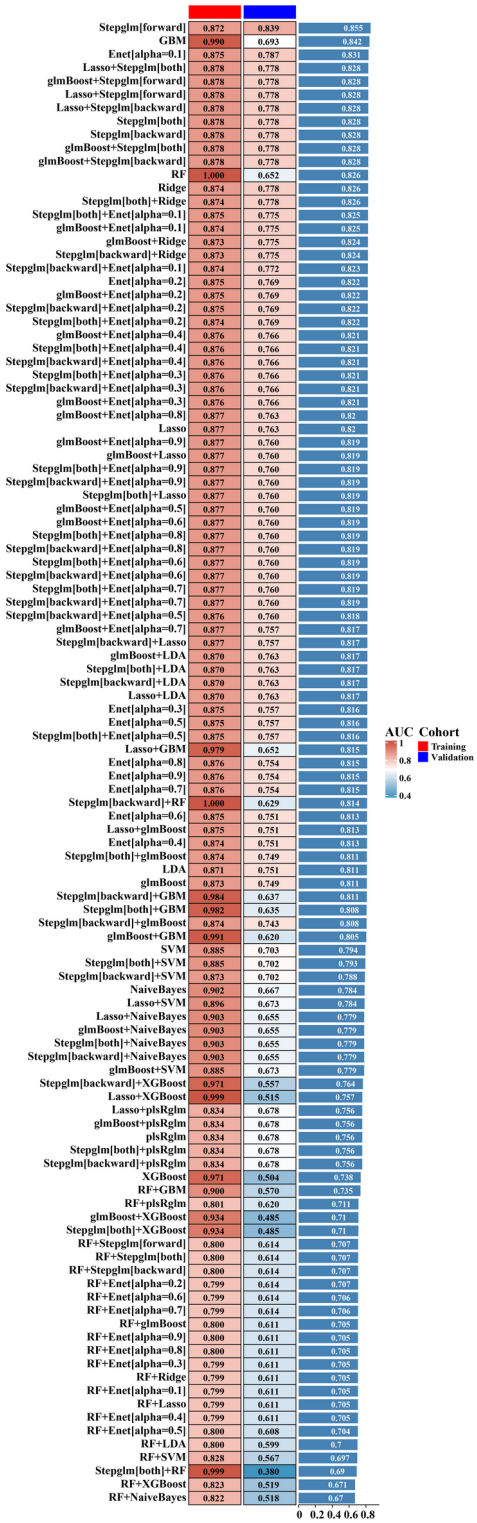

Fig. 4. The pCR prediction model employed the 111 combination machine learning models based on VIT features. VIT_Stepglm [forward] exhibiting the best performance, while AUCs in the training and validation cohorts were 0.872 and 0.839, respectively.

network algorithms. However, when considering radiomics features extracted from chest CT images, a multi-center study indicated that RF performed better, whereas a single-center study found SVM to be superior, among logical regression, RF, and SVM algorithms[35,36]. The emergence of these results is also correlated with feature selection strategies. Thus, whether developing several models can truly represent the best approach for assessing feature suitability is a matter of debate. Against this backdrop, the application of multi-combination models has

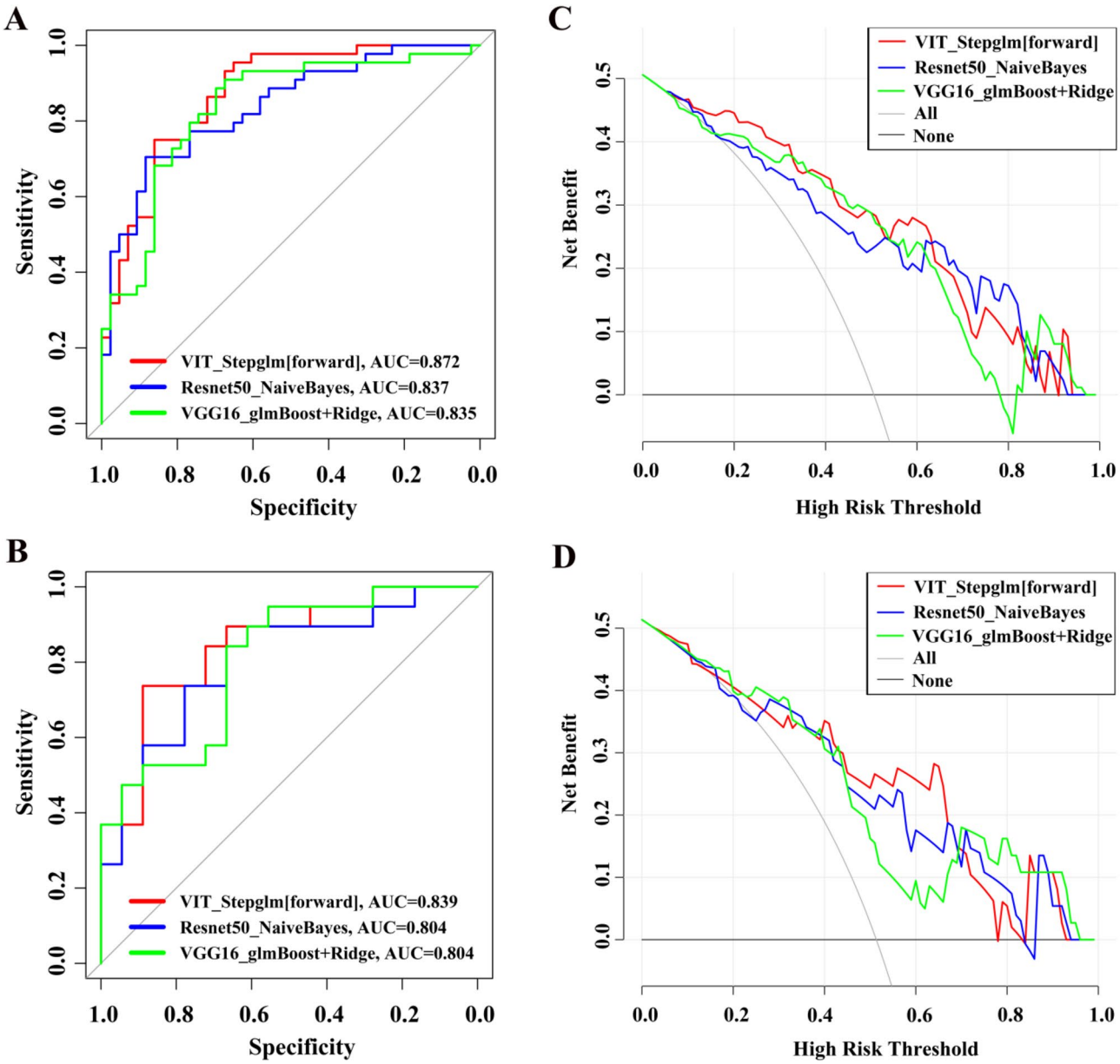

Fig. 5. The optimal algorithm combination results of VIT, ResNet50, and VGG16 features. For VIT, Resnet50, and VGG16 features in the training cohort (**A**) and validation cohort (**B**), the highest AUCs were achieved in the Stepglm [forward], NB, and glmBoost+Ridge combination algorithms, respectively. In the training cohort (**C**) and validation cohort (**D**), DCA demonstrated that the VIT_Stepglm [forward] models possessed good potential clinical application value.

gradually been explored for its potential value. For example, in predicting pancreatic lymph node metastasis, 77 ensemble models were developed based on 10 algorithms, with the Stepglm [backward] algorithm performing the best, achieving an AUC of 0.85[37]. Similarly, in predicting gastric cancer prognosis, 101 mitochondrial-related scoring models were developed based on 12 genes, with the Cox + random survival forest combination yielding the best performance[38].

Against this backdrop, the present study, based on single-center data, first explored the value of VIT features and CNN features in predicting breast cancer NAT-pCR. It was found that VIT features performed better, achieving an AUC of 0.839. Notably, the optimal machine learning approaches for VIT, ResNet50, and VGG16 features differed, being Stepglm [forward], NB, and glmBoost + Ridge, respectively. Encouragingly, the Stepglm [forward] model based on CNN features achieved moderate performance in predicting NAT-pCR in the validation cohort, with AUCs of 0.754 and 0.787, respectively. After reviewing numerous similar publications, it was discovered that the Stepglm algorithm actually performed exceptionally well in many feature selection and prediction tasks, which may be attributed to the algorithm itself. In numerous studies that employed various algorithms for feature selection and model development, for instance, the plsRglm prediction model

Features	Algorithms	Training cohort			Validation cohort		
		AUC	SEN	SPE	AUC	SEN	SPE
VIT	Stepglm[forward]	0.872	0.75	0.86	0.839	0.74	0.89
	Lasso	0.877	0.93	0.70	0.763	0.68	0.89
	Ridge	0.874	0.95	0.70	0.778	0.74	0.78
	Enet[alpha = 0.1]	0.875	0.93	0.70	0.787	0.63	0.89
	Lasso + SVM	0.896	0.89	0.88	0.703	0.68	0.72
	Lasso + glmBoost	0.875	0.93	0.70	0.751	0.58	0.89
	glmBoost + LDA	0.870	0.89	0.74	0.763	0.74	0.78
	Lasso + plsRglm	0.834	0.70	0.84	0.678	0.74	0.61
	RF	1.000	1.00	1.00	0.652	0.32	1.00
	GBM	0.990	0.91	0.98	0.693	0.63	0.72
	Stepglm[backward] + XGBoost	0.971	0.84	0.95	0.557	0.89	0.44
	NaiveBayes	0.902	0.84	0.81	0.667	0.53	0.83
Resnet50	NaiveBayes	0.837	0.70	0.88	0.804	0.74	0.78
	Ridge	0.812	0.73	0.77	0.772	0.68	0.78
	LDA	0.808	0.80	0.70	0.769	0.68	0.78
	glmBoost + Enet[alpha = 0.1]	0.812	0.80	0.72	0.763	0.68	0.78
	Stepglm[forward]	0.812	0.64	0.86	0.754	0.95	0.50
	glmBoost	0.803	0.73	0.74	0.737	0.68	0.72
	SVM	0.840	0.80	0.88	0.735	0.53	0.94
	glmBoost + GBM	0.949	0.89	0.91	0.734	0.95	0.50
	glmBoost + Lasso	0.805	0.55	0.93	0.734	0.68	0.72
	plsRglm	0.774	0.64	0.77	0.646	0.63	0.72
	XGBoost	0.885	0.70	0.93	0.585	0.74	0.50
VGG16	glmBoost + Ridge	0.835	0.89	0.70	0.804	0.84	0.67
	LDA	0.832	0.80	0.79	0.795	0.95	0.56
	Enet[alpha = 0.1]	0.835	0.91	0.67	0.789	0.95	0.56
	Stepglm[forward]	0.832	0.91	0.67	0.787	0.95	0.56
	Lasso + GBM	0.951	0.93	0.91	0.775	0.63	0.83
	glmBoost	0.833	0.84	0.74	0.772	0.95	0.56
	glmBoost + Lasso	0.835	0.86	0.74	0.769	0.95	0.56
	SVM	0.862	0.86	0.86	0.705	0.63	0.78
	Stepglm[both] + NaiveBayes	0.828	0.91	0.67	0.678	0.58	0.78
	Stepglm[both] + XGBoost	0.891	0.75	0.93	0.601	0.47	0.72

Table 2. Development and validation of multi-combination machine learning models.

developed subsequent to feature selection by the Stepglm algorithm was found to be superior in predicting endometriosis compared to other algorithms[39]. Furthermore, multicenter data indicated that the Stepglm model developed after feature selection by the LASSO algorithm was more advantageous in predicting drug-resistant epilepsy[40]. Despite the variation in the most suitable machine learning algorithms for different DL features, there exist algorithms that could excel across multiple DL feature sets. From a feature utilization perspective, the DL_103 feature was utilized a total of 111 times in model development. Through correlation analysis of radiomics features, it was discovered that the DL_103 feature correlated with GLCM ($r = 0.21$) and GLSZM ($r = 0.20$) features, indicating its representation of the complexity of lesion texture distribution, which aligned with the enhanced image weighting by the attention mechanism of VIT. Additionally, when the aforementioned top three machine learning algorithms were employed to develop predictive models based on radiomics features, it was found that the Ridge algorithm performed the best, followed by NB, with Stepglm [forward] performing the worst. However, even the optimal Ridge model (AUC = 0.722) was inferior to several dozen ensemble models based on DL features.

Limitations

Despite the relatively good performance of VIT features in predicting NAT-pCR compared to other CNN algorithms, this study still had certain limitations. Firstly, it was a retrospective study with a small sample size from a single center, and further validation with multi-center large-scale data will be needed to confirm the clinical applicability and robustness of the models. Secondly, the resolution of ultrasound imaging was limited by individual patient differences, such as breast density type and adipose thickness, as well as the acquisition parameters set by physicians. Although the Combat method was used to remove inter-instrument differences, it was still difficult to assess whether different parameters would affect the effectiveness of models. Thirdly, VIT features were extracted based on pre-trained model parameters, which still required large-scale data for model

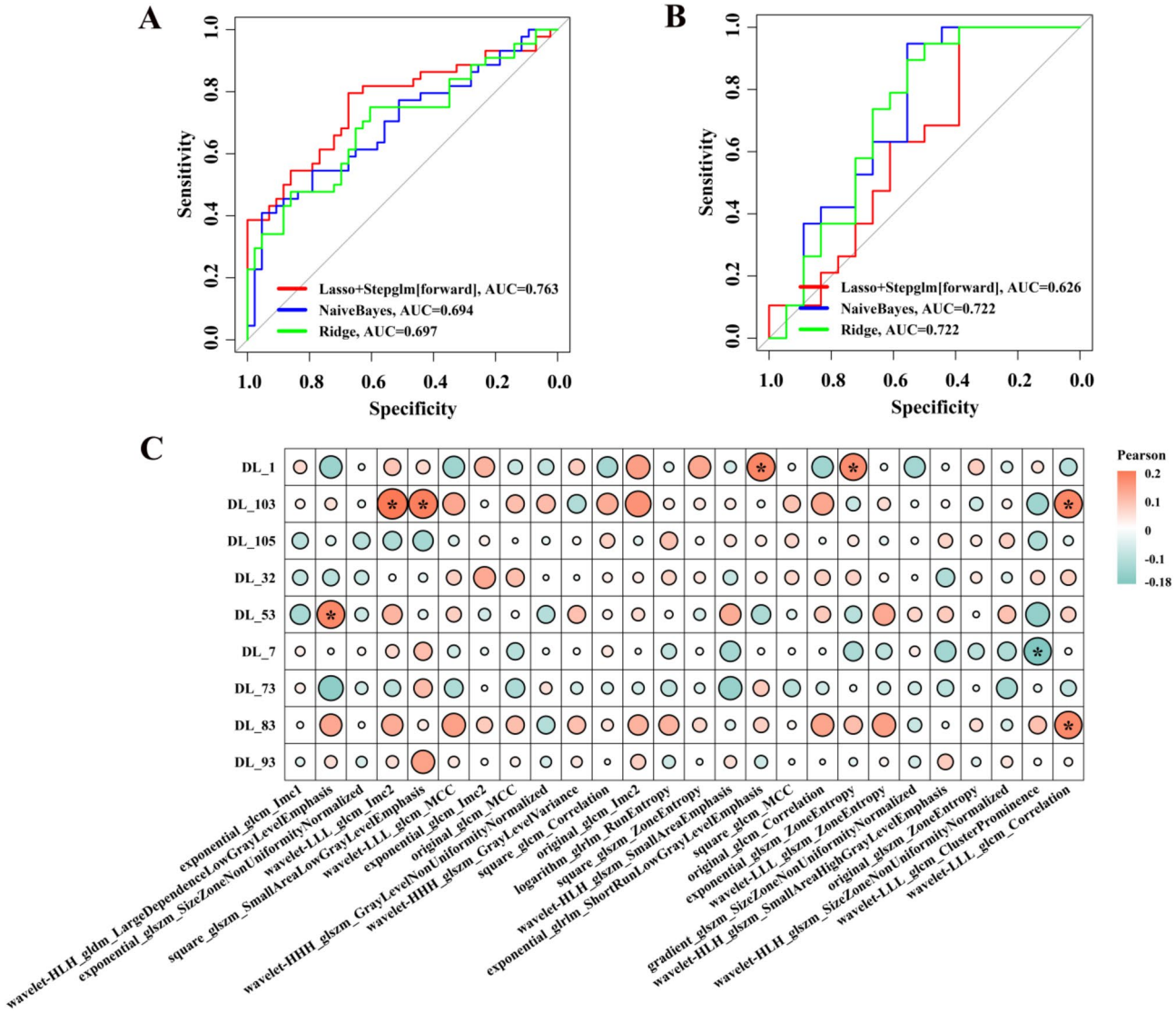

Fig. 6. Interpreting potential clinical significance represented by deep learning features based on radiomics. The predictive models developed based on 26 high-value radiomics features, Ridge algorithm performed the best, while AUCs in the training and validation cohorts were 0.697 (**A**) and 0.722 (**B**), respectively. (**C**) Correlation analysis between 9 VIT features and 26 radiomics features. It was discovered that the DL_103 feature correlated with GLCM (r=0.21) and GLSZM (r=0.20) features, indicating its representation of the complexity of lesion texture distribution, which aligned with the enhanced image weighting by the attention mechanism of VIT.

training and VIT feature extraction. Finally, while attempts had been made to interpret the clinical significance of VIT features based on radiomics features, causal relationships still need to be validated through molecular-level analysis.

Conclusions

In conclusion, most of the machine learning models developed based on VIT features extracted from ultrasound images have shown satisfactory performance. Compared to other CNN features, VIT features exhibit superior performance. However, different machine learning models utilize distinct DL features, resulting in variations in their effectiveness. Consequently, the application of multi-combined models can potentially lead to the discovery of more suitable prediction models. While VIT features hold potential application value, their clinical interpretability still requires further exploration at the molecular level.

Models	Author	Year	*N*	TP	FP	FN	TN
DL_US	Yu FH et al-T	2023	420	139	37	19	225
	Yu FH et al-V	2023	183	51	16	15	101
	Huang JX et al.	2023	255	91	40	14	110
	Gu J et al-T	2024	127	42	16	5	64
	Gu J et al-V	2024	43	11	5	5	22
	Wu L et al-T	2022	242	65	65	6	106
	Wu L et al-V1	2022	197	44	20	13	120
	Wu L et al-V2	2022	212	63	24	14	111
	Wu L et al-V3	2022	150	31	15	7	97
	Taleghamar H et al-V	2022	50	37	3	3	7
	Liu Y et al-T	2022	215	53	5	2	155
	Liu Y et al-V1	2022	95	35	7	6	47
	Liu Y et al-V2	2022	83	17	6	1	59
	Byra M et al.	2021	39	17	6	3	13
RAD_US	Yang M et al-T	2022	152	65	23	24	40
	Yang M et al-V	2022	65	30	9	8	18
	Huang JX et al.	2023	255	87	42	18	108
	Li ZY et al-T	2024	785	252	143	41	349
	Li ZY et al-V	2024	337	110	69	24	134
	Liu J et al-T	2024	324	42	31	42	209
	Liu J et al-V	2024	140	21	14	19	86
	Zhang J et al-T	2023	155	54	57	7	37
	Zhang J et al-V	2023	56	18	8	3	27
DL_PA	Saednia K et al-V	2023	63	14	8	2	39
	Duanmu H et al.	2021	75	30	5	13	27
	Zeng H et al-T	2024	261	59	63	16	123
	Zeng H et al-V1	2024	107	30	25	9	43
	Zeng H et al-V2	2024	72	8	24	2	38
	Li F et al-T	2021	433	52	28	29	324
	Li F et al-V	2021	107	7	1	14	85
	Zhang J et al-T	2023	155	40	17	21	77
	Zhang J et al-V	2023	56	11	3	10	32

Table 3. Literature review on artificial intelligence based on ultrasound or pathology imaging. Note: DL, deep learning; RAD, radiomics; US, ultrasound; PA, pathological.

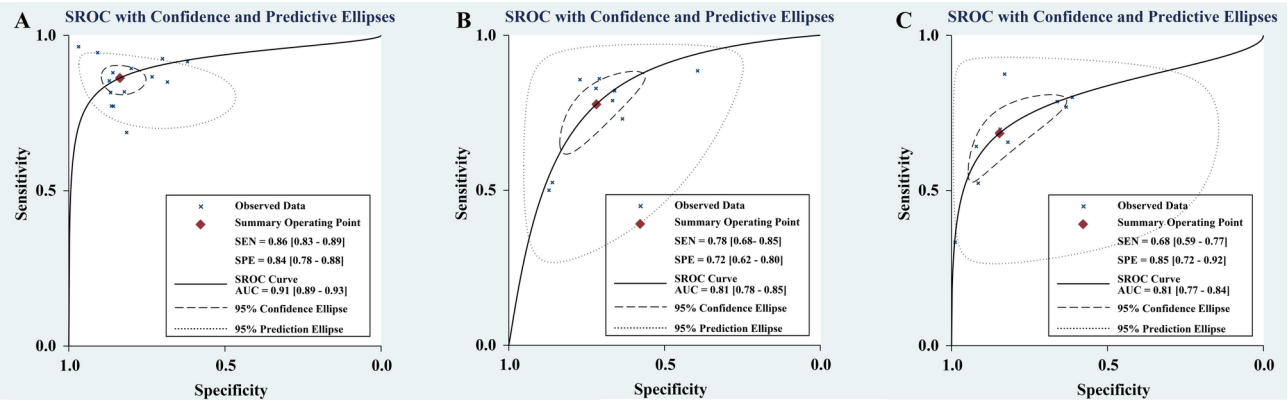

Fig. 7. Literature review on artificial intelligence based on ultrasound or pathology imaging. The meta-analysis results showed that DL features based on ultrasound images (AUC=0.91) were superior to radiomics features based on ultrasound images (AUC=0.81) and DL features based on pathological images (AUC=0.81).

Data availability

The data sets generated and/or analyzed by the current study can be obtained from the corresponding author upon reasonable request.

Received: 15 November 2024; Accepted: 21 April 2025
Published online: 31 December 2025

References

1. Chan, Y. H. Y. et al. Preoperative considerations and benefits of neoadjuvant chemotherapy: insights from a 12-year review of the Hong Kong breast Cancer registry. *Hong Kong Med. J.* **29** (3), 198–207 (2023).
2. Xu, X. et al. The residual cancer burden index as a valid prognostic indicator in breast cancer after neoadjuvant chemotherapy. *BMC Cancer*. **24** (1), 13 (2024).
3. Taleghamar, H. et al. Deep learning of quantitative ultrasound multi-parametric images at pre-treatment to predict breast cancer response to chemotherapy. *Sci. Rep.* **12** (1), 2244 (2022).
4. Gu, J. et al. Deep learning of multimodal ultrasound: stratifying the response to neoadjuvant chemotherapy in breast Cancer before treatment. *Oncologist* **29** (2), e187–e197 (2024).
5. Oh, S. J. et al. Relationship between background parenchymal enhancement on breast MRI and pathological tumor response in breast cancer patients receiving neoadjuvant chemotherapy. *Br. J. Radiol.* **91** (1088), 20170550 (2018).
6. Sui, L. et al. Ultrasound and clinicopathological characteristics-based model for prediction of pathologic response to neoadjuvant chemotherapy in HER2-positive breast cancer: a case-control study. *Breast Cancer Res. Treat.* **202** (1), 45–55 (2023).
7. Jia, Y. et al. Ultrasound features combined with tumor-infiltrating lymphocytes for prediction of pathological response to neoadjuvant chemotherapy in breast cancer. *Med. Ultrason.* **25** (2), 131–138 (2023).
8. Kim, Y. et al. Early prediction of response to neoadjuvant chemotherapy using dynamic Contrast-Enhanced MRI and ultrasound in breast Cancer. *Korean J. Radiol.* **19** (4), 682–691 (2018).
9. Jiang, X. et al. Deep learning for medical Image-Based Cancer diagnosis. *Cancers (Basel)*. **15** (14), 3608 (2023).
10. Cheng, N. et al. Deep Learning-Based classification of hepatocellular nodular lesions on Whole-Slide histopathologic images. *Gastroenterology* **162** (7), 1948–1961e7 (2022).
11. Kumaran, S. Y. & Jeya, J. J. Explainable lung cancer classification with ensemble transfer learning of VGG16, Resnet50 and InceptionV3 using grad-cam. *BMC Med. Imaging*. **24** (1), 176 (2024).
12. Yu, F. H. et al. Pretreatment ultrasound-based deep learning radiomics model for the early prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Eur. Radiol.* **33** (8), 5634–5644 (2023).
13. Zeng, H. et al. Deep learning-based predictive model for pathological complete response to neoadjuvant chemotherapy in breast cancer from biopsy pathological images: a multicenter study. *Front. Physiol.* **15**, 1279982 (2024).
14. Fanizzi, A. et al. Comparison between vision Transformers and convolutional neural networks to predict non-small lung cancer recurrence. *Sci. Rep.* **13** (1), 20605 (2023).
15. Moslemi, A. et al. Apriori prediction of chemotherapy response in locally advanced breast cancer patients using CT imaging and deep learning: transformer versus transfer learning. *Front. Oncol.* **14**, 1359148 (2024).
16. Jiang, M. et al. Ultrasound-based deep learning radiomics in the assessment of pathological complete response to neoadjuvant chemotherapy in locally advanced breast cancer. *Eur. J. Cancer*. **147**, 95–105 (2021).
17. Huang, J. X. et al. Deep learning model based on Dual-Modal ultrasound and molecular data for predicting response to neoadjuvant chemotherapy in breast Cancer. *Acad. Radiol.* **30** (Suppl 2), S50–S61 (2023).
18. Wu, L. et al. An integrated deep learning model for the prediction of pathological complete response to neoadjuvant chemotherapy with serial ultrasonography in breast cancer patients: a multicentre, retrospective study. *Breast Cancer Res.* **24** (1), 81 (2022).
19. Liu, Y. et al. Early prediction of treatment response to neoadjuvant chemotherapy based on longitudinal ultrasound images of HER2-positive breast cancer patients by Siamese multi-task network: A multicentre, retrospective cohort study. *EClinicalMedicine* **52**, 101562 (2022).
20. Byra, M. et al. Early prediction of response to neoadjuvant chemotherapy in breast Cancer sonography using Siamese convolutional neural networks. *IEEE J. Biomed. Health Inf.* **25** (3), 797–805 (2021).
21. Yang, M. et al. Treatment response prediction using Ultrasound-Based Pre-, Post-Early, and Delta radiomics in neoadjuvant chemotherapy in breast Cancer. *Front. Oncol.* **12**, 748008 (2022).
22. Li, Z. Y. et al. Ultrasound-based radiomics-clinical nomogram for noninvasive prediction of residual cancer burden grading in breast cancer. *J. Clin. Ultrasound*. **52** (5), 566–574 (2024).
23. Liu, J. et al. An ultrasound-based nomogram model in the assessment of pathological complete response of neoadjuvant chemotherapy in breast cancer. *Front. Oncol.* **14**, 1285511 (2024).
24. Saednia, K., Tran, W. T. & Sadeghi-Naini, A. A hierarchical self-attention-guided deep learning framework to predict breast cancer response to chemotherapy using pre-treatment tumor biopsies. *Med. Phys.* **50** (12), 7852–7864 (2023).
25. Duanmu, H. et al. Spatial Attention-Based deep learning system for breast Cancer pathological complete response prediction with serial histopathology images in multiple stains. *Med. Image Comput. Comput. Assist. Interv.* **12908**, 550–560 (2021).
26. Li, F. et al. Deep learning-based predictive biomarker of pathological complete response to neoadjuvant chemotherapy from histological images in breast cancer. *J. Transl Med.* **19** (1), 348 (2021).
27. Zhang, J. et al. Development and validation of a radiopathomic model for predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer patients. *BMC Cancer*. **23** (1), 431 (2023).
28. Pesapane, F. et al. Prediction of the pathological response to neoadjuvant chemotherapy in breast Cancer patients with MRI-Radiomics: A systematic review and Meta-analysis. *Curr. Probl. Cancer*. **46** (5), 100883 (2022).
29. O'Donnell, J. P. M. et al. The accuracy of breast MRI radiomic methodologies in predicting pathological complete response to neoadjuvant chemotherapy: A systematic review and network meta-analysis. *Eur. J. Radiol.* **157**, 110561 (2022).
30. Caracciolo, M. et al. Comparison of MRI vs. [18F]FDG PET/CT for treatment response evaluation of primary breast Cancer after neoadjuvant chemotherapy: literature review and future perspectives. *J. Clin. Med.* **12** (16), 5355 (2023).
31. Liang, X., Yu, X. & Gao, T. Machine learning with magnetic resonance imaging for prediction of response to neoadjuvant chemotherapy in breast cancer: A systematic review and meta-analysis. *Eur. J. Radiol.* **150**, 110247 (2022).
32. Li, Z. et al. Ultrasound-based radiomics for early predicting response to neoadjuvant chemotherapy in patients with breast cancer: a systematic review with meta-analysis. *Radiol. Med.* **129** (6), 934–944 (2024).
33. Oğuz, Ç. & Yağanoğlu, M. Detection of COVID-19 using deep learning techniques and classification methods. *Inf. Process. Manag.* **59** (5), 103025 (2022).
34. Latif, G. et al. Novel coronavirus and common pneumonia detection from CT scans using deep Learning-Based extracted features. *Viruses* **14** (8), 1667 (2022).
35. Feng, Y. et al. Prediction of EGFR mutation status in Non-Small cell lung Cancer based on ensemble learning. *Front. Pharmacol.* **13**, 897597 (2022).
36. Liu, Y. et al. Development and validation of machine learning models to predict epidermal growth factor receptor mutation in Non-Small cell lung cancer: A Multi-Center retrospective radiomics study. *Cancer Control*. **29**, 10732748221092926 (2022).

37. Tang, Y. et al. Radiogenomic analysis for predicting lymph node metastasis and molecular annotation of radiomic features in pancreatic cancer. *J. Transl Med.* **22** (1), 690 (2024).
38. Ma, Y. et al. Integrated multi-omics analysis and machine learning developed a prognostic model based on mitochondrial function in a large multicenter cohort for gastric Cancer. *J. Transl Med.* **22** (1), 381 (2024).
39. Zhang, H. et al. Machine learning-based integrated identification of predictive combined diagnostic biomarkers for endometriosis. *Front. Genet.* **14**, 1290036 (2023).
40. Zhang, J. et al. A comprehensive prediction model of drug-refractory epilepsy based on combined clinical-EEG microstate features. *Ther. Adv. Neurol. Disord.* **17**, 17562864241276202 (2024).
41. Wang, H. et al. The predictive value of systemic immune-inflammatory markers before and after treatment for pathological complete response in patients undergoing neoadjuvant therapy for breast cancer: a retrospective study of 1994 patients. *Clin. Transl Oncol.* **26** (6), 1467–1479 (2024).
42. Hagenaars, S. C. et al. Tumor-stroma ratio is associated with Miller-Payne score and pathological response to neoadjuvant chemotherapy in HER2-negative early breast cancer. *Int. J. Cancer.* **149** (5), 1181–1188 (2021).
43. Leithner, D. et al. Impact of combat harmonization on PET Radiomics-Based tissue classification: A Dual-Center PET/MRI and PET/CT study. *J. Nucl. Med.* **63** (10), 1611–1616 (2022).
44. Wu, L. et al. Ultrasound-based deep learning radiomics nomogram for differentiating mass mastitis from invasive breast cancer. *BMC Med. Imaging.* **24** (1), 189 (2024).
45. Zhang, S. & Yuan, G. C. Deep transfer learning for COVID-19 detection and lesion recognition using chest CT images. *Comput. Math. Methods Med.* **2022**, 4509394 (2022).
46. Nojima, S. et al. Deep Learning-Based differential diagnosis of follicular thyroid tumors using histopathological images. *Mod. Pathol.* **36** (11), 100296 (2023).
47. Qin, Q. et al. Development and validation of a multi-modal ultrasomics model to predict response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. *BMC Med. Imaging.* **24** (1), 65 (2024).

Author contributions

LYW, SHL, and FC contributed to the study design, data analysis and process, data acquisition, data interpretation, drafting of manuscript; SHL and CF contributed to the data analysis and process; CJW contributed to the data acquisition and interpretation of endoscopy; SFW and YL contributed to the data acquisition; LYW provided imaging interpretation, evaluation. DYW, and XHX contributed to the study design, critical review of manuscript, and study supervision. LYW, SHL, FC, CJW, YL, DYW, and XHX revised the manuscript additionally. All authors read and approved the final manuscript.

Funding

The authors declared that financial support had been provided to cover the cost of the editorial services for this review. This study was funded by the Science and Technology Planning Project of Maoming City (2024109) and the Maoming Municipal Science and Technology Innovation Development Plan Project (2024kjcxLX075).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.W. or X.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.