Original Article

Artificial intelligence-based classification of breast nodules: a quantitative morphological analysis of ultrasound images

Hao Pan[1]^, Changbei Shi[2], Yuxing Zhang[1,3], Zijian Zhong[1]

[1]School of Electronic Information, Xijing University, Xi'an, China; [2]Department of Nuclear Medicine, Shaanxi Provincial Cancer Hospital, Xi'an, China; [3]School of Medicine, Xijing University, Xi'an, China

Contributions: (I) Conception and design: H Pan, C Shi; (II) Administrative support: Y Zhang; (III) Provision of study materials or patients: H Pan, Z Zhong; (IV) Collection and assembly of data: H Pan, Z Zhong; (V) Data analysis and interpretation: H Pan, Z Zhong; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Yuxing Zhang, MM. School of Electronic Information, Xijing University, Xi'an, China; School of Medicine, Xijing University, No. 1 Xijing Road, Chang'an District, Xi'an 710123, China. Email: zhangyuxing@xijingedu.cn

Background: Accurate classification of breast nodules into benign and malignant types is critical for the successful treatment of breast cancer. Traditional methods rely on subjective interpretation, which can potentially lead to diagnostic errors. Artificial intelligence (AI)-based methods using the quantitative morphological analysis of ultrasound images have been explored for the automated and reliable classification of breast cancer. This study aimed to investigate the effectiveness of AI-based approaches for improving diagnostic accuracy and patient outcomes.

Methods: In this study, a quantitative analysis approach was adopted, with a focus on five critical features for evaluation: degree of boundary regularity, clarity of boundaries, echo intensity, and uniformity of echoes. Furthermore, the classification results were assessed using five machine learning methods: logistic regression (LR), support vector machine (SVM), decision tree (DT), naive Bayes, and K-nearest neighbor (KNN). Based on these assessments, a multifeature combined prediction model was established.

Results: We evaluated the performance of our classification model by quantifying various features of the ultrasound images and using the area under the receiver operating characteristic (ROC) curve (AUC). The moment of inertia achieved an AUC value of 0.793, while the variance and mean of breast nodule areas achieved AUC values of 0.725 and 0.772, respectively. The convexity and concavity achieved AUC values of 0.988 and 0.987, respectively. Additionally, we conducted a joint analysis of multiple features after normalization, achieving a recall value of 0.98, which surpasses most medical evaluation indexes on the market. To ensure experimental rigor, we conducted cross-validation experiments, which yielded no significant differences among the classifiers under 5-, 8-, and 10-fold cross-validation ($P>0.05$).

Conclusions: The quantitative analysis can accurately differentiate between benign and malignant breast nodules.

Keywords: Quantitative analysis; breast ultrasound; artificial intelligence (AI); machine learning

Submitted Nov 22, 2023. Accepted for publication Mar 25, 2024. Published online Apr 26, 2024.
doi: 10.21037/qims-23-1652
View this article at: https://dx.doi.org/10.21037/qims-23-1652

^ ORCID: 0009-0002-4813-9709.

Introduction

Breast cancer is a common malignant tumor in women and can also occur in men. When abnormal cell growth occurs within the breast, it forms a lump or tumor, resulting in breast cancer. According to the American Cancer Society, breast cancer accounts for 29% of all new malignant tumors in women (1) and is the second leading cause of cancer-related death in women (2). In line with this, Global Cancer Statistics 2020 indicated that breast cancer has surpassed lung cancer to become the most common cancer in women and one of the leading causes of cancer-related death among women (3). Fortunately, early detection of lumps can improve treatment outcomes and reduce the mortality rate of breast cancer (4). Therefore, breast examinations are essential for women.

Breast examinations typically include various methods such as mammography, ultrasound, and magnetic resonance imaging (MRI). For women with dense breasts, the detection accuracy of tumor screening with breast X-ray examinations is severely limited. Supplementing X-ray with ultrasound examinations in the United States has the potential to detect early breast cancer that has been missed by breast X-ray examinations, further reducing mortality rates (5-7). Thus, ultrasound in the United States has become a valuable technology for breast cancer screening and differential diagnosis. Ultrasound is noninvasive, cost-effective, and uses nonionizing radiation, making it particularly suitable for detecting tumors in dense breasts and as a result, a useful supplemental tool to mammography (8). Additionally, breast ultrasound can provide high-resolution images, enabling physicians to clearly observe breast structures and pathological features, thereby achieving more accurate diagnoses. However, the grading of the clinical Breast Imaging-Reporting and Data System (BI-RADS) mainly relies on the experience and skills of ultrasound physicians, leading to differences in grading assessments among across different levels of experience, which is not conducive to the standardization of clinical diagnosis (9,10). Moreover, even experienced physicians may make misdiagnoses under high-intensity work pressure (11).

In addressing the issues of limited clinical experience and variations in grading assessments across operators with different levels of experience, artificial intelligence (AI)-based breast cancer detection has been proposed and widely adopted. Over a decade ago, the first computer-aided diagnosis (CAD) systems were proposed to assist radiologists in the interpretation of breast ultrasound exams (12). AI can use deep learning algorithms to analyze and learn from a large number of breast ultrasound images, interpret structural observations and pathological features accurately, and assist physicians in diagnosis (13,14). This can further enable physicians to quickly and accurately determine the nature of tumors, thereby enhancing diagnostic efficiency, providing more accurate diagnostic results and aiding in the detection of early breast cancer lesions that might have gone unnoticed.

In addition to standardizing and objectifying clinical BI-RADS grading, AI algorithms can also intelligently classify and assess breast ultrasound images, reducing grading discrepancies across operators with varying levels of experience (15). This improves diagnostic consistency and makes it possible to use both manual analysis by experienced physicians and CAD as a reference. However, the application of deep learning-based AI algorithms in breast cancer detection presents some challenges. First, the complex parameter-tuning process of deep learning models requires a large number of experiments and optimizations to identify the best model configuration, which may consume significant time and computational resources, limiting the algorithm's rapid application in clinical practice. Second, deep learning model training typically requires a large amount of annotated data, particularly for medical imaging data such as in breast ultrasound images, making the annotation process more challenging and time-consuming (16-18). To address these challenges, this paper proposes a CAD method based on morphological features that quantify physicians' evaluation criteria for the benign or malignant nature of breast nodules. This can assists physicians in diagnosis and reduce the rate of misdiagnosis.

In this study, we attempted to analyze the morphological features of breast nodules by quantitatively analyzing the regularity of the margins, the clarity of the margins, and the internal echoes. We further used machine learning methods, including logistic regression (LR) and decision tree (DT) to establish a quantitative analysis model, providing a basis for the quantitative analysis in the classification of nodules into benign or malignant types. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-1652/rc).

Table 1 The performance of the normalized classifier

Machine learning models	Min-max scaling	Standardization	Decimal scaling	None
LR	92.06%	95.23%	79.36%	76.19%
SVM	95.22%	96.03%	97.6%	74.6%

LR, logistic regression; SVM, support vector machine.

Methods

Data preprocessing

AI image recognition training typically involves four steps: database creation, interest identification, image feature extraction and analysis, and prediction model building. In this study, a dataset was used (19) to develop a prediction mode. The dataset was collected at Baheya Hospital from January to December 2018 and consists of 600 female patients. The ages of these patients range from 25 to 75 years, and their lesions vary in size, making them highly suitable for our study. Furthermore, the dataset comprises 487 benign images and 210 malignant images. In this dataset, there are a few instances in which a single image contains two tumors, but the corresponding masks are separated into two different images. This discrepancy could have affected our numerical calculations, so these images were excluded. After excluding some images unsuitable for this experiment, a total of 422 benign images and 210 malignant images were ultimately used. We aimed to calculate breast nodule morphological features but recognize the potential for the model to deteriorate due to scale and dimensional differences, and thus preprocessing of the calculated data was necessary. Normalization is a widely used technique for image recognition. By eliminating dimensional differences between features or variables, we can ensure that their impact on the model is fair and consistent. This process improves the stability and accuracy of the model and makes the weights of different morphological features more reasonable and interpretable. Common normalization methods include the following:

(I) Min-Max scaling: this method maps the data in a linear manner within the range of the minimum and maximum values provided according to the following formula:

$$X_{normalized} = \frac{(X - X_{\min})}{(X_{\max} - X_{\min})} \quad [1]$$

This method rescales the data to fit within the range of values between 0 and 1.

(II) Standardization: this method calculates the mean and standard deviation and transforms the data into a distribution with a mean of 0 and a standard deviation of 1 as follows:

$$X_{normalized} = \frac{(X - \mu)}{\sigma} \quad [2]$$

The variables μ and σ represent the mean and standard deviation of the entire sample. This method can bring the distribution of data closer to a standard normal distribution.

(III) Decimal scaling: this method transforms the data into fixed intervals such as [–1, 1] or [–0.5, 0.5] via the division of the maximum absolute value in the data as follows:

$$X_{normalized} = \frac{X}{10j} \quad [3]$$

In the method of decimal scaling normalization, "k" is usually determined based on the maximum absolute value of the dataset. Its purpose is to scale the data to a range smaller than 1, facilitating subsequent processing.

Table 1 displays the impact of these three normalization methods on the LR and support vector machine (SVM) algorithms. By analyzing the table provided, we can observe that various classifiers display varying levels of sensitivity toward the original features. However, once normalization is introduced, it brings all the features within the same range, ultimately improving the classification ability of the classifier.

Morphological features

In determining whether a breast nodule is benign or malignant, the characteristics of its margin and echo are crucial indicators. In this study, we quantified these features using Python code (Python Software Foundation) and assessed their impact on the nature of the nodule. Custom-built algorithms implemented in Python were used to calculate morphological features. We used receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC) to evaluate the significance of these features.

When examining boundaries in breast ultrasound examinations using AI and machine learning, it is important to focus on their clarity and regularity. Clear boundaries are a significant morphological feature and can aid in determining whether a nodule is benign or malignant.

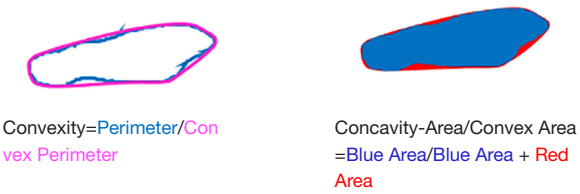

Figure 1 Convexity and concavity.

The clarity and regularity of boundaries are important features analyzed using AI algorithms. Benign nodules have clear, regular boundaries (circumscribed as per the fifth edition of the BI-RADS), typically presenting as round or oval-shaped (oval shape as per the BI-RADS), whereas the boundaries of malignant nodules are usually irregular (not circumscribed; i.e., indistinct, angular, spiculated, and microlobulated as per the BI-RADS). Additionally, echo patterns are an important feature of AI algorithms, with significant differences between benign and malignant nodules. Typically, benign nodules present as hyperechoic and isoechoic (as per the BI-RADS), while malignant nodules usually appear hypoechoic (as per the BI-RADS) and exhibit heterogeneous internal echoes (complex echoic as per the BI-RADS). Therefore, all of these factors should be analyzed in the assessment of nodules.

In this study, the analysis of edge and echo features incorporated the following five formulas, all of which demonstrated excellent classification performance:

$$\text{Convexity} = \frac{\text{P}}{\text{Convex Perimeter}} \quad [4]$$

$$\text{Concavity} = \frac{\text{Area}}{\text{Convex Area}} \quad [5]$$

$$\text{Inertia Moment} = \sum_i \sum_j (i-j)^2 P(i,j|d,\theta) \quad [6]$$

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{N} \quad [7]$$

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \quad [8]$$

Hamyoon *et al.* proposed the first and second equations to assess the regularity of breast nodules (20). The application of these formulas is based on the significant differences in margin regularity between benign and malignant nodules. Benign nodules typically have more regular margins, while malignant nodules often exhibit irregular and jagged features. By applying these two formulas, we can effectively distinguish between benign and malignant nodules with different degrees of margin regularity. The specific calculation methods are shown in *Figure 1*.

The third formula calculates the inertia moment near the tumor area, which is a measure of the local changes in the image and the distribution of matrix values, reflecting the clarity of the image and the degree of texture grooves. Since benign nodules typically have clear boundaries, they can distinguish the tumor area from the surrounding tissue, while malignant nodules have relatively vague boundaries, making it difficult to distinguish the boundary and surrounding tissue. Therefore, the numerical difference in inertial moment can effectively distinguish between benign and malignant nodules.

The fourth and fifth formulas are used to calculate the mean and variance of the pixel values in the nodule region, which can reflect the distribution of grayscale values within the area. In ultrasound images, the brightness of echoes is associated with tissue density and acoustic impedance, and different types of tumors or normal tissues exhibit distinct echo characteristics. Benign nodules typically exhibit hypoechoic or isoechoic echoes, while malignant nodules often manifest as hyperechoic or mixed echoes. Thus, by calculating the mean and variance of the nodule region, it becomes possible to effectively differentiate between benign and malignant nodules. The specific process of our study is shown in *Figure 2*.

Indicators and classification for evaluation

In this study, various measures were used to evaluate the performance of classifiers, including accuracy, precision, recall, F1-score, ROC curve, and AUC. These metrics are often used to assess the accuracy of classifiers. Accuracy, precision, recall, and F1-score are common evaluation metrics in machine learning and can comprehensively assess the quality of classification results from different perspectives. Before these metrics are calculated, it is necessary to understand the concepts of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). True positives are samples that are predicted as positive and indeed are positive, FN are samples that are predicted as negative but are actually positive, FP are samples that are predicted as positive but are actually positive, and TN are samples that are predicted as negative and are indeed negative. The specific formulas for these metrics are as follows:

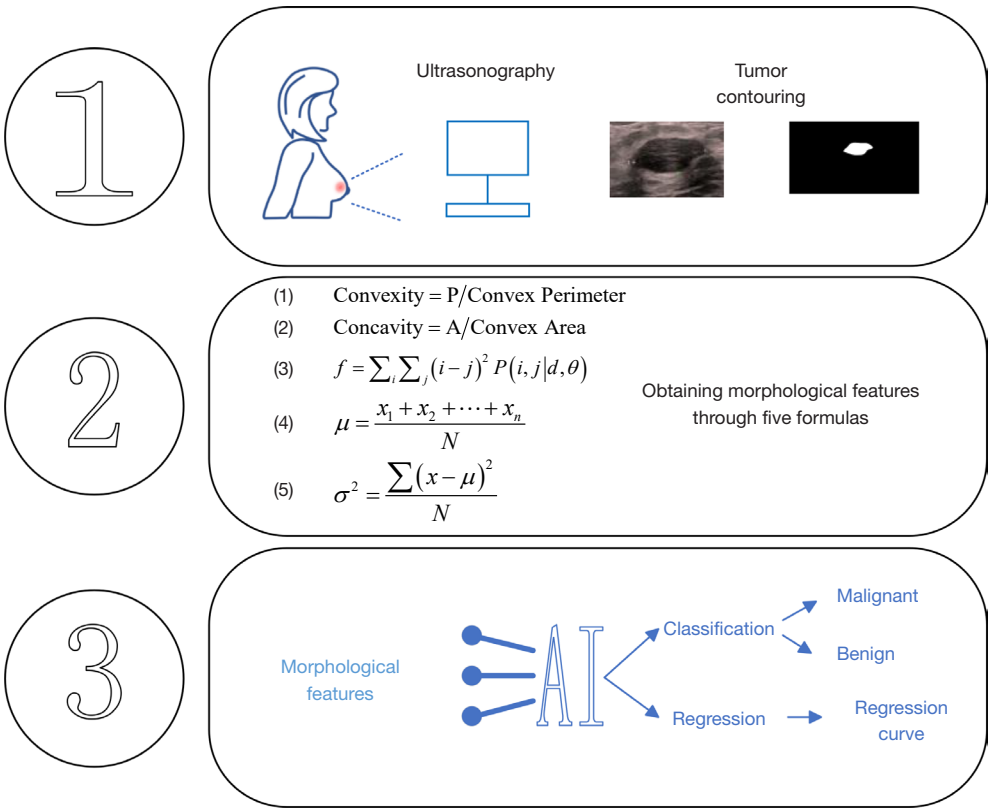

Figure 2 Experimental process.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad [9]$$

$$\text{Recall} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad [10]$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad [11]$$

$$\text{F1-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad [12]$$

The recall value indicates how many actual positive cases are correctly identified, which is crucial in the medical field. If the recall is low, it means that more cases can be missed, which could have serious consequences, such as delayed treatment, disease progression, or even death. Therefore, it is essential to maintain high recall to reduce the risk of missed diagnoses and ensure that patients receive timely and appropriate medical care.

Furthermore, our study aimed to go beyond individual analysis of each morphological feature's diagnostic performance and to conduct a comprehensive assessment by incorporating the five equations mentioned earlier. Multiple classifiers were used to identify the most appropriate classification method, resulting in a more precise and compelling diagnostic method that encompassed various perspectives.

Multifeature joint classification

The five formulas mentioned above can provide insights into the distinctive characteristics of breast nodule edges and echoes. However, it is not sufficient to rely solely on individual features to analyze the benign or malignant nature of breast nodules, as this may lack rigor. Therefore, it is necessary to employ various techniques such as LR, SVM, DT, and other algorithms to incorporate multiple features for analysis. These algorithms use feature matrices to store the feature values of samples, and each feature must undergo standardization to eliminate dimensional disparities among samples. This allowed us to explore the interrelationships among multiple features while identifying their respective contribution to distinguishing between benign and malignant nodules.

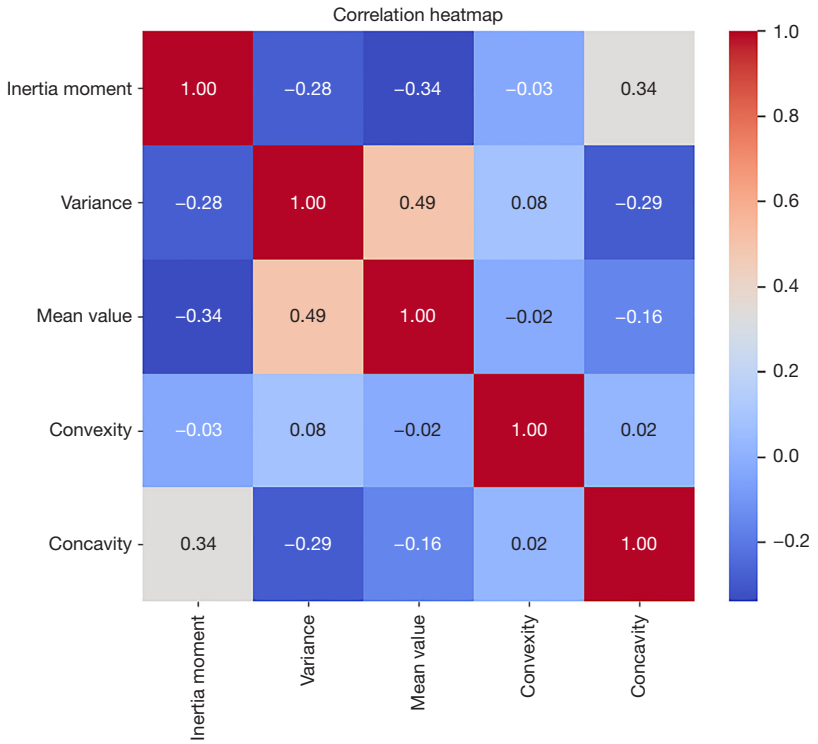

Figure 3 Correlation coefficients.

Table 2 Morphological features

Feature name	Benign	Malignant
Inertia moment	8,109.91 (4,585.76, 14,580.43)	3,127.51 (2,022.28, 5,038.41)
Variance	1.91 (6.74, 3.21)	3.74 (2.31, 6.74)
Mean value	2.23 (10.70, 5.36)	7.19 (4.29, 10.70)
Convexity	1.05±0.01	1.16±0.08
Concavity	0.98±0.01	0.88±0.06

Data are presented as the median (IQR) or as the mean ± standard deviation. IQR, interquartile range.

Multifeature joint prediction model

In addition to joint classification, multiple features can be combined to establish a predictive model. Commonly used predictive models include LR, SVM, etc., which can predict probabilities based on input feature values and form a probability curve. The probability curve can be combined with the 4A (2–10% probability of malignancy), 4B (10–50% probability of malignancy), and 4C (50–95% probability of malignancy) classifications mentioned in the fifth edition of the BI-RADS, assisting in more accurate clinical diagnosis.

Results

Distribution of features among the groups

The correlation coefficients among various features are shown in the *Figure 3*. The correlation coefficient between the variance and mean of the breast mass was 0.49, the coefficient between concavity and Inertia moment was 0.34, and the coefficient between the inertia moment and mean value was −0.34. These findings suggest that there may be some redundant information among these features. The correlation coefficients of the other features were relatively small, indicating limited redundancy among them.

Individual classification results for each feature

Table 2 presents the numerical values of various morphological features. *Table 3* presents the results of individual feature classification, with the AUC values indicating the quality of the classification results and a higher AUC value corresponding to a better classification

outcome. A significance test was conducted to determine the significant differences between the groups. The results indicated that the features of concavity and concave points demonstrated the best performance in individual classification. This suggests significant differences in the regularity of boundaries between benign and malignant nodules. However, relying solely on a single feature for classification is not sufficiently rigorous, and multiple data points should be integrated for classification purposes.

Results of joint classification using multiple features

The results of multiple classifiers for the joint classification of various features are presented in *Table 4*. Five common classifiers were selected for this analysis: LR, naive Bayes, K-nearest neighbor (KNN), DT, and SVM. The performance parameters of these classifiers were recorded, and the data were normalized using the maximum and minimum values before being compared with the original results. Before normalization, except for the DT classifier, the performance of the other classifiers was not ideal. Due to the characteristics of its algorithm, the DT could achieve good classification results even without normalization. After the maximum and minimum normalization was applied, the performance indicators of all classifiers improved significantly.

Table 3 AUC and P values of the morphological features

Feature name	AUC	P value
Inertia moment	0.793	0.03
Variance	0.725	0.03
Mean value	0.772	0.02
Convexity	0.988	0.008
Concavity	0.987	0.005

AUC, area under the curve.

Cross-validation

To ensure the experiment's accuracy, we performed a cross-validation test. The results of this test are presented in *Tables 5*-7. The purpose of this test was to evaluate the effectiveness of the two models on various data subsets. We used the K-fold cross-validation method, in which divided the dataset into K equal-sized subsets and used each subset as the test data while the other subsets served as the training data. Based on the cross-validation results, we concluded that the classification algorithms used in the 5-, 8-, and 10-fold cross-validation tests were not significantly different from one another (P>0.05).

Comparison with traditional diagnostic methods

To validate the effectiveness of the experimental results, a comparison was made between the research findings of this study and those of traditional diagnostic methods. Aristokli *et al.* summarized the sensitivity and specificity

Table 4 The machine learning metrics for multifeature joint classification

Classification	AUC	Accuracy	Precision	Recall	F1-score
LR	0.873 (0.798, 0.949)	0.823 (0.757, 0.889)	0.838 (0.794, 0.883)	0.911 (0.843, 0.979)	0.871 (0.820, 0.923)
Min-max scaling + LR	0.993 (0.986, 0.999)	0.944 (0.926, 0.962)	0.928 (0.908, 0.948)	0.992 (0.986, 0.999)	0.959 (0.944, 0.975)
SVM	0.767 (0.682, 0.853)	0.748 (0.679, 0.817)	0.813 (0.764, 0.861)	0.811 (0.733, 0.889)	0.809 (0.751, 0.867)
Min-max scaling + SVM	0.996 (0.991,0.999)	0.976 (0.964, 0.988)	0.975 (0.966, 0.987)	0.988 (0.979,0.996)	0.982 (0.971,0.992)
KNN	0.705 (0.630, 0.779)	0.685 (0.632, 0.737)	0.761 (0.727, 0.796)	0.771 (0.697, 0.845)	0.763 (0.714, 0.811)
Min-max scaling + KNN	0.997 (0.994,0.999)	0.971 (0.958, 0.984)	0.969 (0.956, 0.983)	0.988 (0.979, 0.996)	0.978 (0.967, 0.990)
NB	0.820 (0.764. 0.877)	0.756 (0.694, 0.818)	0.805 (0.766, 0.845)	0.835 (0.758, 0.912)	0.818 (0.762, 0.873)
Min-max scaling + NB	0.986 (0.977, 0.995)	0.947 (0.930, 0.964)	0.977 (0.966, 0.989)	0.942 (0.924, 0.961)	0.960 (0.944, 0.975)
DT	0.952 (0.919, 0.984)	0.965 (0.943, 0.986)	0.965 (0.933, 0.996)	0.976 (0.965, 0.987)	0.973 (0.958, 0.987)
Min-max scaling + DT	0.967 (0.953, 0.999)	0.945 (0.923, 0.967)	0.934 (0.910, 0.958)	0.987 (0.975, 0.999)	0.960 (0.948, 0.972)

The data in parenthesis are presented as the 95% confidence interval. AUC, area under the curve; LR, logistic regression; SVM, support vector machine; KNN, K-nearest neighbor; NB, naive Bayes; DT, decision tree.

Table 5 The results of fivefold cross-validation

Machine learning models (K=5)	AUC	Accuracy	Precision	Recall	F1-score
LR	0.994 (0.991, 0.997)	0.938 (0.901, 0.974)	0.923 (0.877, 0.970)	0.992 (0.987, 0.997)	0.956 (0.931, 0.981)
SVM	0.987 (0.982, 0.992)	0.936 (0.894, 0.978)	0.941 (0.892, 0.990)	0.969 (0.955, 0.983)	0.953 (0.927, 0.980)
KNN	0.974 (0.960, 0.988)	0.926 (0.893, 0.959)	0.938 (0.903, 0.973)	0.957 (0.929, 0.985)	0.946 (0.936, 0.976)
NB	0.967 (0.952, 0.982)	0.923 (0.892, 0.954)	0.977 (0.947, 0.997)	0.924 (0.869, 0.979)	0.938 (0.907, 0.969)
DT	0.935 (0.914, 0.956)	0.958 (0.933, 0.983)	0.963 (0.931, 0.994)	0.973 (0.962, 0.984)	0.969 (0.953, 0.984)

The data in parenthesis are presented as the 95% confidence interval. AUC, area under the curve; LR, logistic regression; SVM, support vector machine; KNN, K-nearest neighbor; NB, naive Bayes; DT, decision tree.

Table 6 The results of eightfold cross-validation

Machine learning models (K=8)	AUC	Accuracy	Precision	Recall	F1-score
LR	0.996 (0.991, 0.999)	0.966 (0.943, 0.989)	0.969 (0.940, 0.998)	0.983 (0.975, 0.991)	0.975 (0.959, 0.991)
SVM	0.994 (0.989, 0.999)	0.934 (0.892, 0.976)	0.927 (0.878, 0.976)	0.988 (0.974, 0.999)	0.954 (0.928, 0.981)
KNN	0.986 (0.972, 0.999)	0.949 (0.916, 0.982)	0.961 (0.923, 0.999)	0.966 (0.941, 0.991)	0.962 (0.939, 0.985)
NB	0.985 (0.971, 0.999)	0.939 (0.904, 0.974)	0.977 (0.955, 0.999)	0.933 (0.878, 0.987)	0.952 (0.922, 0.981)
DT	0.952 (0.924, 0.979)	0.958 (0.933, 0.983)	0.965 (0.933, 0.996)	0.976 (0.965, 0.987)	0.968 (0.952, 0.985)

The data in parenthesis are presented as the 95% confidence interval. AUC, area under the curve; LR, logistic regression; SVM, support vector machine; KNN, K-nearest neighbor; NB, naive Bayes; DT, decision tree.

Table 7 The results of 10-fold cross-validation

Machine learning models (K=10)	AUC	Accuracy	Precision	Recall	F1-score
LR	0.983 (0.987, 0.998)	0.939 (0.898, 0.981)	0.927 (0.882, 0.971)	0.995 (0.989, 0.999)	0.958 (0.931, 0.984)
SVM	0.997 (0.994, 0.999)	0.966 (0.943, 0.989)	0.969 (0.941, 0.998)	0.983 (0.968, 0.998)	0.975 (0.959, 0.991)
KNN	0.987 (0.978, 0.997)	0.952 (0.928, 0.975)	0.959 (0.928, 0.990)	0.973 (0.955, 0.991)	0.965 (0.949, 0.981)
NB	0.986 (0.976, 0.997)	0.945 (0.921, 0.969)	0.979 (0.959, 0.998)	0.940 (0.907, 0.974)	0.958 (0.939, 0.976)
DT	0.956 (0.922, 0.990)	0.958 (0.933, 0.983)	0.968 (0.936, 0.999)	0.976 (0.961, 0.990)	0.972 (0.9540.989)

The data in parenthesis are presented as the 95% confidence interval. AUC, area under the curve; LR, logistic regression; SVM, support vector machine; KNN, K-nearest neighbor; NB, naive Bayes; DT, decision tree.

of conventional diagnostic approaches (including mammography, ultrasound, and MRI) based on the published literature, as shown in *Table 8* (21).

The results of multifeature joint prediction

Figure 4 presents the results of multifeature joint prediction using SVM and LR, which were two classifiers with the best overall performance. The additional curve in the graph, labeled as the voting classifier prediction, represents the results obtained by combining the predictions of SVM and LR through a voting mechanism. Voting classifier is an ensemble learning technique that combines the predictions of multiple classifiers to improve the overall classification performance. The x-axis in *Figure 4* represents the normalized values, while the y-axis represents the

Table 8 Comparison of the study results with those of conventional diagnosis

Diagnostic method	Sensitivity (%)	Specificity (%)
MRI	94.6	74.2
MM	54.5	85.5
US	67.2	76.8
CE-MRI	91.5	64.7
CE-MM	90.5	52.6
CE-MM + US	90.5	61.4
MRI + MM	95.8	70.1
MRI + US	92.3	76.8
MM + US	72.2	87.8
MRI + MM + US	97.7	63.3
Proposed model	98	96

MRI, magnetic resonance imaging; MM, mammography; US, ultrasound; CE-MM, contrast-enhanced mammography; CE-MRI, contrast-enhanced MRI.

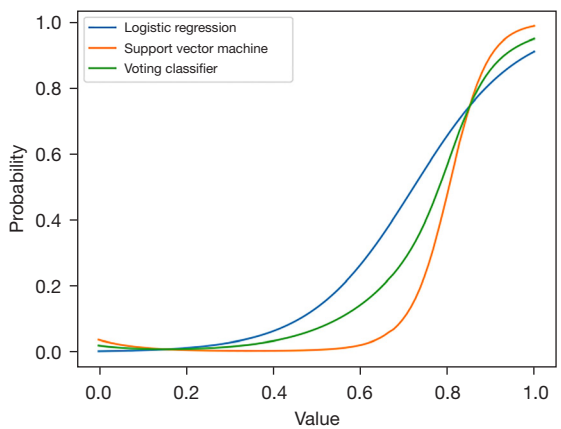

Figure 4 Prediction curve. The x-axis represents the normalized values, while the y-axis represents the probability of the corresponding value being malignant.

probability of the corresponding value being malignant.

Discussion

This study introduces a method for analyzing the morphology of breast ultrasound images for classification and prediction. The rapid and efficient nature of ultrasound examinations makes this approach potentially more accurate and valuable. While existing and proposed solutions for breast imaging mostly rely on mammography, further research into ultrasound imaging in this field is essential, as there is little evidence available for its use in breast imaging (22).

Additionally, in developing research protocols, it is crucial to evaluate and consider robustness, replicability, and external validation (23). This paper presents an AI method based on morphological features, which differs from previous studies that solely compare deep learning models with doctors' diagnostic results (24). An AI approach based on morphological features is more convenient, eliminating the need for complex parameter adjustments associated with deep learning and the facilitation of replication (25). A texture-combined SVM classifier was developed that incorporate five morphological features: overlap ratio, aspect ratio, circularity, normalized residual value (NRV), and P ratio (ratio of convex hull perimeter to tumor perimeter). When these features were combined with texture features and input into the SVM algorithm, the accuracy rate reached 95.83%. Pereira *et al.* extracted seven features and used linear discriminant analysis (LDA) to identify the best features for distinguishing between benign and malignant lesions. Their experiments showed that circularity and NRV were the most relevant features for breast lesion classification. They also demonstrated significant correlations between the overlap ratio, NRV, circularity, and area ratio, suggesting that the information contained in other features used for classification might be redundant (26). In contrast, for our study, the correlation coefficients of the five extracted morphological features were low, indicating that the correlation between them was weak and that the majority of these five features are not redundant. Unlike previous studies, which typically applied features from the computer vision field to breast nodules, our study focused more on analyzing the characteristics of breast nodules themselves, quantifying the indicators used by doctors to evaluate the benignity and malignancy of breast nodules, which more closely aligns with the methods doctors use to diagnose breast nodules.

Based on the classification results of individual features, convexity and concavity performed well when used alone. This suggests a strong relationship between the regularity of the breast nodule boundary and its benign or malignant nature. Compared to that of convexity and concavity, the classification performance of the other features was slightly worse. This may be due to the fact that a single feature cannot fully reflect the complex characteristics of a breast nodule, making it necessary to use multiple features for

classification. It is worth noting that the classification based on individual features still has certain reference value, as it can be used for preliminary screening or rapid evaluation, thereby reducing unnecessary pathological biopsies; however, it should not be relied upon as the sole basis for final differentiation.

Selecting the optimal model for breast nodule benign–malignant classification among various classifiers is an important decision that requires considering factors such as classification accuracy, interpretability, and computational complexity. According to the results of multifeature joint classification, the accuracy the classifiers—except for the DT algorithm—for processing the raw data was between 70% and 80%. This was due to the differences in scaling factors among various features, which led to insufficient accuracy in the classifiers. The DT, due to its algorithm's characteristics, achieved good results even when dealing with nonnormalized data. However, after normalization, the accuracy of most classifier algorithms was similar to that of the DT. Among the various evaluation indicators, the recall value (sensitivity), which is most critical for medical diagnosis, reached 98%. Based on *Table 8*, it can be observed that the diagnostic results obtained through AI morphological analysis surpass the majority of conventional diagnostic outcomes. In addition to classification accuracy, this study also proposed using these features to build predictive models that correspond to the BI-RADS grading system, allowing doctors to compare their diagnostic results with those of CAD systems and thus obtain the most accurate diagnosis.

To ensure the validity of our findings, we conducted cross-validation experiments using 5-, 8-, and 10-fold methods. None of classifiers used in the experiment showed significant differences (P>0.05) under cross-validation. With regard to all the K values, LR and SVM performed the best, with high AUC values, accuracy, precision, recall, and F1-score. This indicates that these two classifiers have excellent performance for this problem. KKN exhibited a relatively stable with performance, with high AUC values and accuracy under different K values, but its precision and recall decreased slightly. This suggests that the KNN model may miss some positive samples, which can hinder diagnostic efficacy.

In this study, we conducted a quantitative analysis of the classification of breast nodules into benign and malignant categories. We developed a predictive model that can perform quantitative analysis and identify various features that are associated with the classification of breast nodules into benign and malignant categories. After conducting experiments with multiple classifiers, the accuracy of classification was highly satisfactory, indicating the potential of this model as an auxiliary tool for clinical application. However, some limitations in this study should be mentioned. Breast nodules consist of a diverse group of pathological conditions, including various benign and malignant subgroups. Unfortunately, it is challenging to obtain pathological images of some rare diseases, which limited our study to common benign and malignant lesions. Additionally, our study used a dataset comprising 422 benign images and 210 malignant images. Although these images are sufficient for simple classification tasks, they may be inadequate for complex regression tasks, leading to imprecise prediction curves. We also did not compare clinical decisions in our study. Thus, there are several areas for improvement in future research: First, we need to extract morphological features of rare diseases to enable our model to adapt to the majority of diseases. Second, we need to expand our dataset by including a sufficient number of images to verify whether the prediction curve changes. Third, we should consider collaborating with clinical experts to validate and evaluate our model. We can then include the decisions made by experienced clinicians of varying levels and compare them with those of the quantitative analysis model, thereby enhancing the reliability and applicability of our classifier.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-1652/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-1652/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.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin 2016;66:7-30.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. 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.
4. Lee CH. Screening mammography: proven benefit, continued controversy. Radiol Clin North Am 2002;40:395-407.
5. Gordon PB, Goldenberg SL. Malignant breast masses detected only by ultrasound. A retrospective review. Cancer 1995;76:626-30.
6. Kolb TM, Lichy J, Newhouse JH. Occult cancer in women with dense breasts: detection with screening US--diagnostic yield and tumor characteristics. Radiology 1998;207:191-9.
7. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. Radiology 2002;225:165-75.
8. Anderson BO, Shyyan R, Eniu A, Smith RA, Yip CH, Bese NS, Chow LW, Masood S, Ramsey SD, Carlson RW. Breast cancer in limited-resource countries: an overview of the Breast Health Global Initiative 2005 guidelines. Breast J 2006;12 Suppl 1:S3-15.
9. Youk JH, Jung I, Yoon JH, Kim SH, Kim YM, Lee EH, Jeong SH, Kim MJ. Comparison of Inter-Observer Variability and Diagnostic Performance of the Fifth Edition of BI-RADS for Breast Ultrasound of Static versus Video Images. Ultrasound Med Biol 2016;42:2083-8.
10. Schwab F, Redling K, Siebert M, Schötzau A, Schoenenberger CA, Zanetti-Dällenbach R. Inter- and Intra-Observer Agreement in Ultrasound BI-RADS Classification and Real-Time Elastography Tsukuba Score Assessment of Breast Lesions. Ultrasound Med Biol 2016;42:2622-9.
11. Giess CS, Frost EP, Birdwell RL. Difficulties and errors in diagnosis of breast neoplasms. Semin Ultrasound CT MR 2012;33:288-99.
12. Chen DR, Hsiao YH. Computer-aided Diagnosis in Breast Ultrasound. J Med Ultrasound 2008;16:46-56.
13. Gao Y, Geras KJ, Lewin AA, Moy L. New Frontiers: An Update on Comput-er-Aided Diagnosis for Breast Imaging in the Age of Artificial Intelligence. AJR Am J Roentgenol 2019:212:300-7.
14. Fujioka T, Mori M, Kubota K, Oyama J, Yamaga E, Yashima Y, Katsuta L, Nomura K, Nara M, Oda G, Nakagawa T, Kitazume Y, Tateishi U. The Utility of Deep Learning in Breast Ultrasonic Imaging: A Review. Diagnostics (Basel) 2020;10:1055.
15. Becker AS, Mueller M, Stoffel E, Marcon M, Ghafoor S, Boss A. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. Br J Radiol 2018;91:20170576.
16. Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Shen D, Chen CM. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Sci Rep 2016;6:24454.
17. Fleury E, Marcomini K. Performance of machine learning software to classify breast lesions using BI-RADS radiomic features on ultrasound images. Eur Radiol Exp 2019;3:34.
18. Cao Z, Duan L, Yang G, Yue T, Chen Q. An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. BMC Med Imaging 2019;19:51.
19. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief 2019;28:104863.
20. Hamyoon H, Yee Chan W, Mohammadi A, Yusuf Kuzan T, Mirza-Aghazadeh-Attari M, Leong WL, Murzoglu Altintoprak K, Vijayananthan A, Rahmat K, Ab Mumin N, Sam Leong S, Ejtehadifar S, Faeghi F, Abolghasemi J, Ciaccio EJ, Rajendra Acharya U, Abbasian Ardakani A. Artificial intelligence, BI-RADS evaluation and morphometry: A novel combination to diagnose breast

cancer using ultrasonography, results from multi-center cohorts. Eur J Radiol 2022;157:110591.
21. Aristokli N, Polycarpou I, Themistocleous SC, Sophocleous D, Mamais I. Comparison of the diagnostic performance of Magnetic Resonance Imaging (MRI), ultrasound and mammography for detection of breast cancer based on tumor type, breast density and patient's history: A review. Radiography (Lond) 2022;28:848-56.
22. Kim J, Kim HJ, Kim C, Kim WH. Artificial intelligence in breast ultrasonography. Ultrasonography 2021;40:183-90.
23. Martin-Noguerol T, Luna A. External validation of AI algorithms in breast radiology: the last healthcare security checkpoint? Quant Imaging Med Surg 2021;11:2888-92.
24. Jiang M, Li CL, Luo XM, Chuan ZR, Lv WZ, Li X, Cui XW, Dietrich CF. Ultrasound-based deep learning radiomics in the assessment of pathological complete response to neoadjuvant chemotherapy in locally advanced breast cancer. Eur J Cancer 2021;147:95-105.
25. Prabusankarlal KM, Thirumoorthy P, Manavalan R. Assessment of combined textural and morphological features for diagnosis of breast masses in ultrasound. Hum Cent Comput Inf Sci 2015;5:1-17.
26. Pereira WC, Alvarenga AV, Infantosi AF, Macrini L, Pedreira CE. A non-linear morphometric feature selection approach for breast tumor contour from ultrasonic images. Comput Biol Med 2010;40:912-8.

Cite this article as: Pan H, Shi C, Zhang Y, Zhong Z. Artificial intelligence-based classification of breast nodules: a quantitative morphological analysis of ultrasound images. Quant Imaging Med Surg 2024;14(5):3381-3392. doi: 10.21037/qims-23-1652

