European Journal of Radiology 114 (2019) 175–184

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Research article

Prediction of molecular subtypes of breast cancer using BI-RADS features based on a “white box” machine learning approach in a multi-modal imaging setting

Mingxiang Wu[a], Xiaoling Zhong[a], Quanzhou Peng[b], Mei Xu[a], Shelei Huang[a], Jialin Yuan[a], Jie Ma[a,*], Tao Tan[c,*]

[a] Department of Radiology, Shenzhen People’s Hospital, No.1017 Dongmen North Road, Luohu District, Shenzhen, Guangdong, 518020, PR China
[b] Department of Pathology, Shenzhen People’s Hospital, No.1017 Dongmen North Road, Luohu District, Shenzhen, Guangdong, 518020, PR China
[c] Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands

ARTICLE INFO

Keywords:
Breast cancer
Molecular subtype
MRI
Mammography
Decision tree
Machine learning

ABSTRACT

Purpose: To develop and validate an interpretable and repeatable machine learning model approach to predict molecular subtypes of breast cancer from clinical metainformation together with mammography and MRI images.

Methods: We retrospectively assessed 363 breast cancer cases (Luminal A 151, Luminal B 96, HER2 76, and BLBC 40). Eighty-two features defined in the BI-RADS lexicon were visually described. A decision tree model with the Chi-squared automatic interaction detector (CHAID) algorithm was applied for feature selection and classification. A 10-fold cross-validation was performed to investigate the performance (i.e., accuracy, positive predictive value, sensitivity, and F1-score) of the decision tree model.

Results: Seven of the 82 variables were derived from the decision tree-based feature selection and used as features for the classification of molecular subtypes including mass margin calcification on mammography, mass margin types of kinetic curves in the delayed phase, mass internal enhancement characteristics, non-mass enhancement distribution on MRI, and breastfeeding history. The decision tree model accuracy was 74.1%. For each molecular subtype group, Luminal A achieved a sensitivity, positive predictive value, and F1-score of 79.47%, 75.47%, and 77.42%, respectively; Luminal B showed a sensitivity, positive predictive value, and F1-score of 64.58%, 55.86%, and 59.90%, respectively; HER2 had a sensitivity, positive predictive value, and F1-scores of 81.58%, 95.38%, and 87.94%, respectively; BLBC showed sensitivity, positive predictive value, and F1-scores of 62.50%, 89.29%, and 73.53%, respectively.

Conclusions: We applied a complete “white box” machine learning method to predict the molecular subtype of breast cancer based on the BI-RADS feature description in a multi-modal setting. By combining BI-RADS features in both mammography and MRI, the prediction accuracy is boosted and robust. The proposed method can be easily applied widely regardless of variability of imaging vendors and settings because of the applicability and acceptance of the BI-RADS.

1. Introduction

Breast cancer is one of the leading causes of death among women worldwide [1]. It is a heterogeneous disease with several distinct molecular subtypes based on receptor status and immunochemistry staining, including expression of the estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2–neu (HER2), proliferation marker Ki67, and the epidermal growth factor receptor (EGFR) [2,3]. There are four major molecular subtypes: luminal A, luminal B, human epidermal growth factor receptor-2 overexpressing (HER2), and basal-like breast cancer (BLBC). Each type shows distinct recurrence and survival rates that are the major factors for choosing various therapy techniques [4,5]. For example, the luminal A type shows the best survival rates [6], whereas HER2 and BLBC have the worst survival rates [7] but are more sensitive to preoperative chemotherapy [8,9].

Determining molecular subtype in breast cancer facilitates making a precise diagnosis and personalizing treatment. However, biopsy

[*] Corresponding authors.
E-mail addresses: majie688@hotmail.com (J. Ma), t.tan1@tue.nl (T. Tan).

https://doi.org/10.1016/j.ejrad.2019.03.015
Received 13 January 2019; Received in revised form 15 March 2019; Accepted 19 March 2019
0720-048X/ © 2019 Published by Elsevier B.V.

Table 1
Immuno-phenotype and molecular subtype of breast cancer patients.

Molecular Subtypes	Immuno-phenotype	Number of Cases
Luminal A	ER + and/or PR+, HER2−,CK5/6 ±, and Ki67 < 14%;	151
Luminal B	ER + and/or PR+,CK5/6 ±, HER2+, or Ki67 ≥ 14%; or PR < 20%	96
HER2	ER−, PR−, HER2+, CK5/6 ±	76
BLBC	ER−, PR−, HER2− (triple negative), CK5/6+, and/or EGFR+.	40

sampling is invasive, and molecular analysis requires specialized equipment and technical expertise which are not commonly available in developing countries, thus hindering its application. Mammography and magnetic resonance imaging (MRI) have played an evolving role in the screening, diagnosis, and treatment response evaluation of breast cancer due to their non-invasiveness and common usage in oncology imaging. In mammography, the luminal molecular type is often associated with architectural distortion, the HER2-positive cancers are more likely to be associated with calcifications and can more frequently be multifocal, and the triple negative cancers are most likely to be associated with a mass [10,26]. On MRI, the basal subtype cancers are more frequently round, the HER2 lesions often show a smooth margin, and luminal B subtype lesions normally demonstrate heterogeneous internal enhancement. Therefore, radiological features, such as tumor size, morphology, shape, and enhancement characteristics have the potential to be used for differentiating breast cancer subtypes [11–15].

In recent years, radiomics/radiogenomics studies on the molecular classification of breast cancer have emerged, and these exploit image texture analysis for imaging feature extraction, followed by feature classification using machine learning methods [16–18]. Texture analysis is advantageous for detecting subtle quantitative imaging features that cannot be identified by observation or direct measurements on the images. Nonetheless, quantitative texture analysis used for feature extraction may be hindered by the variability of imaging acquisition protocols (e.g., variations in pulse sequence parameters, contrast agent injection procedures, and imaging coverage, etc.) and image reconstruction methods [19]. This limitation also reduces the reproducibility of experiments and leads to difficulties in making comparisons or performing meta-analyses on the different studies [18].

The generalizability of imaging features defined by radiologists remains stable irrespective of different imaging vendors or protocols while image-processing-based features often vary. However, limited memory capacity in humans [20], which hinders the analysis and inference of high-dimensional characteristic variables, has led to an increased focus on the utility of machine learning as compared to features described by radiologists, such as the BI-RADS lexicon.

The decision tree is one of the most interpretable classification algorithms in machine learning [21], and it can easily be converted into rules to be applied to doctors' daily work. Therefore, combining the feature extraction ability of radiologists with the decision tree algorithm might be helpful for establishing better reproducibility and a more comprehensible model than image-processing-based radiomics.

In the present study, we used the diagnostic index proposed in the common BI-RADS manual (2013 BI-RADS® Atlas, 5th Edition) to derive quantitative variables of clinical characteristics and imaging features. We established a standard feature table across the observer and examination platform by using fuzzy mathematics in an attempt to build a standardized model that can be used for the prediction of molecular subtypes in breast cancer.

2. Materials and methods

2.1. Participants

The local institutional review board approved this retrospective study. Informed consent was obtained from all participants.

We used our institution's radiology information System (RIS) and pathological information system to identify and review data on all breast cancer patients admitted to our hospital between January 2003 and December 2016. The inclusion criteria were as follows: 1) the patient had been diagnosed with breast cancer histopathologically, and the molecular subtype of the breast cancer had been determined; 2) the patient had undergone preoperative mammography and MRI imaging of the breasts within a month and had pathological results available from a surgery or biopsy; 3) the patient's clinical history was available. The exclusion criteria were as follows: 1) the patient only underwent MRI or mammography and not both exams; 2) the clinical data were missing or incomplete; 3) the patients had undergone breast cancer treatment (e.g., chemotherapy or radiation therapy) before mammography or MRI. Ultimately, a total of 363 female patients aged 45 ± 10 years (age range of 21–77 years) were included in this study.

Each of the patients had been diagnosed with one of four different molecular cancer subtypes based on immune histochemical results after surgery or biopsy (Table 1) in accordance with the 13th St. Gallen International Breast Cancer Conference (2013) Expert Panel criteria [22]. All immune histochemical analyses for subtype differentiation were reviewed cooperatively by at least two pathologists.

2.2. Mammography

Digital mammography was performed on Siemens Mammomat Inspiration (Siemens Healthineers, Erlangen, Germany) using mediolateral oblique and craniocaudal views of both breasts. The protocol was as follows: VA10 = 20–40, 40–70 mA s, angle = −45 – 180 degrees, magnetic force = 60–90 N, thickness = 40–70 mm, anode/filter = W/Rh, and glandular dose =0.0 mGy.

2.3. MRI acquisition

Breast MRI was performed on a 1.5 T scanner (Magnetom Avanto, Siemens Healthineers, Erlangen, Germany) or a 3.0 T scanner (Magnetom Skyra, Siemens Healthineers, Erlangen, Germany) with a dedicated breast coil. The patients were placed in the prone position to ensure that the breasts were appropriately fixed in the gantry of the scanner.

For both the 1.5 T and 3.0 T scans, routine T1-weighted (T1W) and T2-weighted (T2W) turbo spin echo (TSE) sequences with and without fat suppression on the transverse plane were acquired initially. Diffusion-weighted images (DWI) were acquired using single-shot echo planar imaging with the following parameters: TR /TE = 6400/97 ms (1.5 T) and 5700 /59 ms (3.0 T), matrix = 192 × 192 (1.5 T) and 340 × 170 (3.0 T), slice thickness = 4 mm, and b-values = 50/500/1000s/mm^2 (1.5 T) and 50/400/800 s/mm^2 (3.0 T). Next, dynamic contrast-enhanced (DCE) imaging was performed during the injection of Gd-DTPA (0.1 mmol/kg) at a dose of 15 ml at a rate of 2.5 ml/s. The duration of the DCE scan was 6 min and 41 s, including one pre-contrast scan and five post-contrast scans in the transverse plane at a time interval of 30 s. The DCE parameters were as follows: FOV = 380 × 380 mm^2, TR/TE = 4.7/1.7 ms, flip angle = 10°, matrix = 448 × 372, and slice thickness/gap = 1.6/0.3 mm. Last, 3D T1W FLASH imaging was performed in the sagittal plane after DCE with the following parameters: FOV = 220 × 220 mm^2, TR/TE = 5.16/2.38 ms

Table 2
The 82 feature values of breast cancer based on the BI-RADS manual.

Features	Description	Value
MAMMOGRAPHY		
Breast Composition	1. Almost entirely fatty tissue	25
	2. Scattered areas of fibroglandular density	0,1
	3. Heterogeneously dense	0,1
	4. Extremely dense	0,1
Mass Shape	5. Oval	0,1
	6. Round	0,1
	7. Irregular	0,1
Mass Margin	8. Circumscribed	0,1
	9. Obscured	0,1
	10. Microlobulated	0,1
	11. Indistinct	0,1
	12. Spiculated	0,1
Mass Density	13. High density	0,1
	14. Equal density	0,1
	15. Low density	0,1
	16. Fat-containing	0,1
Calcification Morphology	17. No calcification	0 or 1
	18. Amorphous	0,1
	19. Coarse heterogeneous	0,1
	20. Fine pleomorphic	0,1
	21. Fine linear or fine linear branching	0,1
Calcification Distribution	22. Diffuse	0,1
	23. Regional	0,1
	24. Grouped	0,1
	25. Linear	0,1
	26. Segmental	0,1
Architectural Distortion	27. No architectural distortion	0 or 1
	28. Architectural disorder	0,1
	29. Architecture disappeared	0,1
MAGNETIC RESONANCE IMAGING		
Amount of Fibroglandular Tissue (FGT) Level	30. Minimal	0,1
	31. Mild	0,1
	32. Moderate	0,1
	33. Marked	0,1
Amount of Fibroglandular Tissue (FGT) Symmetry	34. Symmetric	0,1
	35. Asymmetric	0,1
Mass Shape	36. Oval	0,1
	37. Round	0,1
	38. Irregular	0,1
Mass Margin	39. Circumscribed	0,1
	40. Irregular	0,1
	41. Spiculated	0,1
Mass Internal Enhancement Characteristics	42. No enhancement	0 or 1
	43. Homogeneous	0,1
	44. Heterogeneous	0,1
	45. Rim enhancement	0,1
	46. Dark internal septations	0,1
Mass Size	47. Smeared-out boundary	0
	T1: Size ≤ 2 cm	1
	T2: 2 cm < size ≤ 5 cm	2
	T3: Size > 5 cm	3
Intramammary Lymph Node	48. Normal (< 1 cm)	0
	Lymphadenopathy (≥ 1 cm)	1
Kinetic Curve in Initial Phase	49. Slow	0,1
	50. Medium	0,1
	51. Fast	0,1
Kinetic curve in Delayed phase	52. Persistent	0,1
	53. Plateau	0,1
	54. Washout	0,1
Non-mass Enhancement Distribution	55. Focal	0,1
	56. Linear	0,1
	57. Segmental	0,1
	58. Regional	0,1
	59. Multiple regions	0,1
	60. Diffuse	0,1
Non-mass Internal Enhancement Patterns	61. Homogeneous	0,1
	62. Heterogeneous	0,1
	63. Clumped	0,1
	64. Clustered ring	0,1
MAMMOGRAPHY		
Associated Features	65. Nipple retraction	0 or 1
	66. Nipple invasion	0 or 1
	67. Skin retraction	0 or 1
	68. Skin thickening	0 or 1
	69. Skin invasion	0 or 1
	70. Axillary adenopathy	0 or 1
	71. Pectoralis muscle invasion	0 or 1
	72. Chest wall invasion	0 or 1
	73. Architectural distortion	0 or 1
Clinical Information		
74. Breast Cancer Family History		0 or 1
75. Oral Contraceptive History(≥ 3 months)		0 or 1
76. Reproductive History		0 or 1
77. Breastfeeding History(≥ 1 month)		0 or 1
78. Multiple Abortion History(≥ 2)		0 or 1
79. Breast Prosthesis Implantation		0 or 1
80. Nipple Discharge		0 or 1
81. Skin Abnormality		0 or 1
82. Age		[21,77]

(1.5 T) and 4.66/1.68 ms (3.0 T), flip angle = 25° (1.5 T) and 10° (3.0 T), matrix = 512 × 512 (1.5 T) and 448 × 372 (3.0 T), and slice thickness/gap = 1.0/0.0 mm (1.5 T) and 1.5/0.3 mm (3.0 T).

2.4. *Clinical metainformation*

The collected clinical information included age, breast cancer family history, oral contraceptive history (≥ 3 months), reproductive history, breastfeeding history (≥ 1 month), abortion history, history of nipple discharge, and the presence of skin abnormities.

2.5. *Feature analysis*

Based on the common BI-RADS manual (American College of Radiology, 2013, 5th Edition), a total of 82 features (29 derived from mammography, 44 derived from MRI, and nine based on clinical information) associated with breast cancer were evaluated (Table 2). We used both mammography and MRI features in order to mirror clinical practice in which patients suspected of having breast cancer undergo both types of imaging exams. However, “associated features” cited the items only in the MRI examination because MRI provides better delineation of the lesions due to superior soft tissue contrast compared with mammography.

To solve potential discrepancies between subjective assessments of the imaging features by the radiologists, the fuzzy mathematics method and swarm intelligence theory were used. Swarm intelligence theory holds that the aggregation of people's opinions can lead to more accurate decision-making. The assumption underlying this theory is that all individuals have a common target destination that they wish to reach, but that individuals navigate toward this target with some error. If the average preference over all the group members is taken into account for each single step when moving toward the target, the error with which the group moves toward the target decreases as a nonlinear function of group size [23]. For example, 3-person groups perform better than the best individuals on letters-to-numbers problems [24]. In our study, the images were independently scored by five radiologists, and the mean score was calculated as a probability of each feature. For example, if 3 of 5 (60%) radiologists considered a breast gland to be “round”, whereas the remaining 2 of 5 (40%) radiologists considered it to be “irregular,” the score was round = 0.6 and irregular = 0.4. These score values were continuous variables between 0 and 1, where 0 represented no possibility (i.e., none of the radiologists had a positive opinion), and 1 represented a positive consensus (i.e., all five radiologists had a

Table 3
Characteristics or clinical conditions.

Characteristics or clinical condition	All (N = 363)	Luminal A (n = 151)	Luminal B (n = 96)	HER2 (n = 76)	BLBC (n = 40)	p-value
Age	45.35 ± 10.15	46.60 ± 10.13	48.31 ± 10.92	41.89 ± 7.65	40.13 ± 8.99	0.000*
Positive family history	87 (24.0%)	41 (27.2%)	17 (17.7%)	15 (19.7%)	14 (35.0%)	0.097
No family history	276 (76.0%)	110 (72.8%)	79 (82.3%)	61 (80.3%)	26 (65.0%)	
Oral contraceptive history	51 (14.0%)	17 (11.3%)	11 (11.5%)	14 (18.4%)	9 (22.5%)	0.167
No oral contraceptive history	312 (86.0%)	134 (88.7%)	85 (88.5%)	62 (81.6%)	31 (77.5%)	
Reproductive history	307 (84.6%)	129 (85.4%)	79 (82.3%)	70 (92.1%)	29 (72.5%)	0.310
No reproductive history	56 (15.4%)	22 (14.6%)	17 (17.7%)	6 (7.9%)	11 (27.5%)	
Breastfeeding history	179 (49.3%)	102 (67.5%)	42 (43.8%)	20 (26.3%)	15 (37.5%)	0.000*
No breastfeeding history	184 (50.7%)	49 (32.5%)	54 (56.3%)	56 (73.7%)	25 (62.5%)	
Multiple abortion history	39 (10.7%)	17 (11.3%)	10 (10.4%)	8 (10.5%)	4 (10.0%)	0.994
No multiple abortion history	324 (89.3%)	134 (88.7%)	86 (89.6%)	68 (89.5%)	36 (90.0%)	
Breast prosthesis implantation	19 (5.2%)	8 (5.3%)	5 (5.2%)	4 (5.3%)	2 (5.0%)	1.000
No breast implantation	344 (94.8%)	143 (94.7%)	91 (94.8%)	72 (94.7%)	38 (95.0%)	
Nipple discharge	64 (17.6%)	25 (16.6%)	13 (13.5%)	18 (23.7%)	8 (20.0%)	0.350
No nipple discharge	299 (82.4%)	126 (83.4%)	83 (86.5%)	58 (76.3%)	32 (80.0%)	
Skin abnormalities	59 (16.3%)	27 (17.9%)	12 (12.5%)	12 (15.8%)	8 (20.0%)	0.637
No skin abnormalities	304 (83.7%)	124 (82.1%)	84 (87.5%)	64 (84.2%)	32 (80.0%)	

* $p < 0.05$.

positive opinion). Other imaging features (e.g., “associated features” and “clinical information”) and clinical information except for “age” were scored using binary categorical variables 0 or 1 to avoid discrepancies between the radiologists. The scoring method is shown in Table 2.

2.6. Decision tree

The decision tree is the most interpretable classification algorithm in machine learning [22]. The decision tree is in the form of a tree structure, where each nonterminal node represents a decision on one attribute, each branch reflects the decision output, and each leaf node represents one classification result. Compared with traditional algorithms such as the support vector machine, Naive Bayes, and the increasing depth neural network, each step of the decision tree derivation process is clear and comprehensible and is easily converted into reference rules in clinical practice.

2.7. Statistical analysis

Data were checked for inconsistencies, and the invalid samples were excluded from the analyses. A descriptive analysis was performed to describe the characteristics of the demographic variables. Clinical condition variables were compared between distinct molecular subtypes using one-way ANOVAs.

The original variables were initially analyzed. The principle of analysis was to calculate the Pearson correlation coefficient and p-value of the input and output variables and obtain the importance of each input variable (Importance = 1 – p). The smaller the p-value, the closer the importance to 1, which meant the more reliable the correlation, and the more important the input variable.

The classification model was built using the decision tree due to its ability to handle non-linear features and to account for variable interactions. The Chi-squared automatic interaction detector (CHAID) algorithm was used for the decision tree analysis. Chi-squared for determining node splitting and category merging were calculated using the Pearson method. The maximum tree depth (level below root) was set for three levels to avoid overfitting and to ensure the clarity and interpretability of the results. The parameters used in the decision tree were experimentally set up as follows: Alpha for Splitting = 0.05, Alpha for Merging = 0.05, and Maximum iterations for convergence = 100. Epsilon = 0.001 determines the convergence criterion (i.e., how much change must occur for iterations to continue). The Bonferroni correction was performed to avoid α error accumulation. The rule sets (i.e., distinct paths for classification) were calculated and reported. The distribution characteristics of the final selected features of the four molecular subtypes were demonstrated using bar plots and radar maps. All the analyses were performed using SPSS Clementine 12.0 (SPSS Inc., Chicago, IL, USA). A p-value less than 0.05 was considered statistically significant.

2.8. Model evaluation

To verify the classification accuracy, a 10-fold cross-validation method was used to randomly divide the entire sample into ten groups. In each round, nine of the groups were used as a training set and one served as a validation set. This process was repeated ten times until each group of the sample had been verified, and the mean accuracy, sensitivity, positive predictive value (PPV), and F1-scores of all the training and validation sets were calculated. The confusion matrix between the actual value and the prediction value of all samples was also calculated to make a comprehensive evaluation of the model. The parameters were defined as follows:

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + True\ negative + False\ positive + False\ negative}$$

$$Sensitivity = \frac{True\ positive}{True\ positive + False\ negative}$$

$$PPV = \frac{True\ positive}{True\ positive + False\ positive}$$

$$F1 - score = 2\frac{Recall * Precision}{Recall + Precision}$$

3. Results

3.1. Clinical information for different molecular subtypes

The clinical information for the 363 enrolled patients with different molecular subtypes is summarized in Table 3. One-way ANOVA tests showed significant associations between the molecular subtypes and age ($p < 0.001$) and between molecular subtypes and breastfeeding history ($p < 0.001$). The ages at the onset of HER2 and BLBC subtype patients were lower than the ages at the onset of luminal subtype patients.

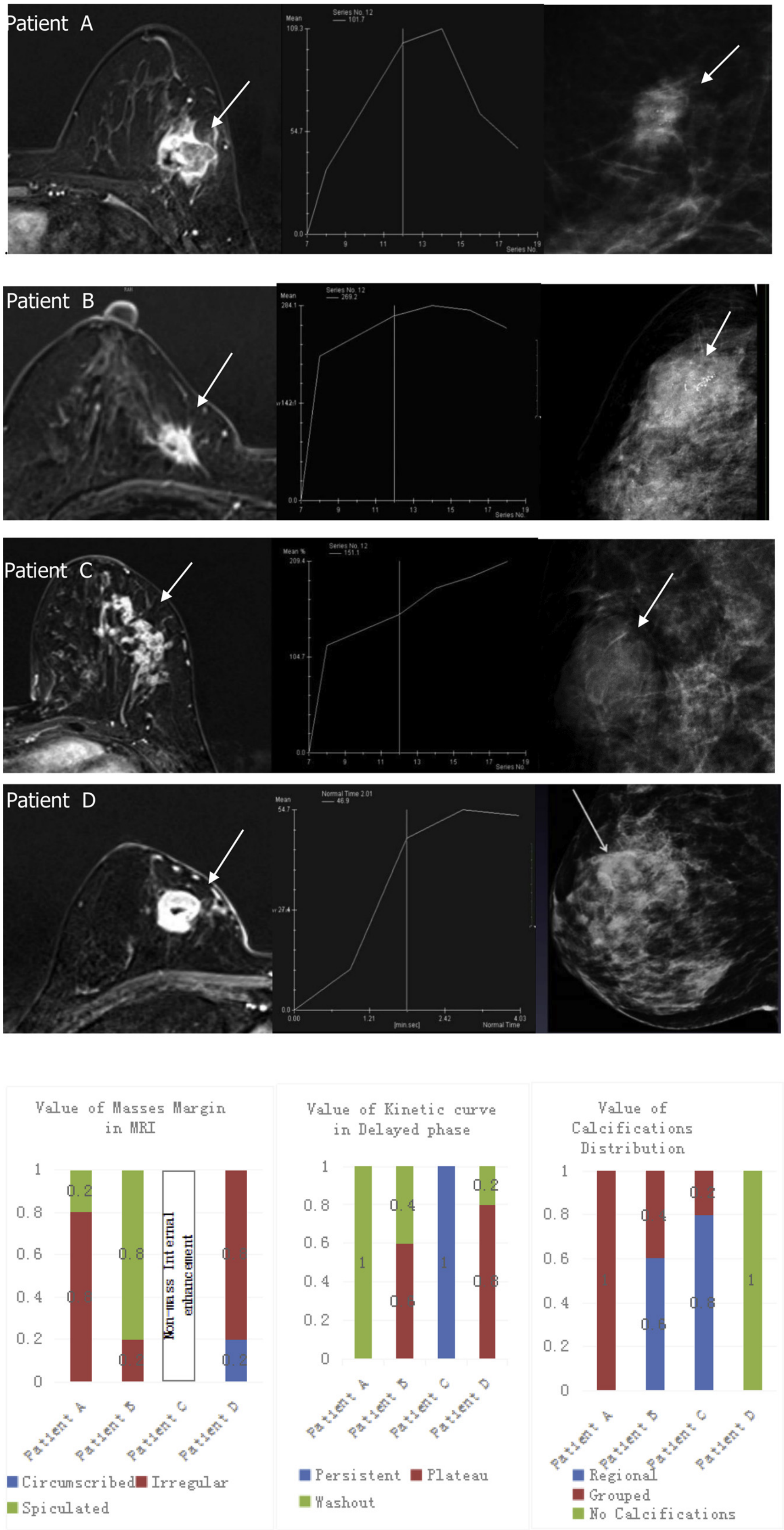

Fig. 1. Each of the upper four rows shows the masses margin on MRI, kinetic curve, and calcifications distribution on mammography images of patient A–D. Patient A: a 45-year-old woman with Luminal A type breast cancer. Patient B: a 45-year-old woman with Luminal B type breast cancer. Patient C: a 45-year-old woman with HER2 type breast cancer. Patient D: a 39-year-old woman with BLBC type breast cancer. The last row shows the scores by 5 radiologists based on the features extracted from the above images. Distribution of the scores of each feature are shown as the bar graphs.

3.2. Feature analysis

The image features of each patient were scored independently by five radiologists, and the average value of each feature was taken as a probability between 0 and 1. Examples of four patients are shown in Fig. 1. Patient A was a 45-year-old woman with Luminal A type breast cancer, and patient B was a 45-year-old woman with Luminal B type breast cancer. Patient C was a 45-year-old woman with HER2 type breast cancer, and patient D was a 39-year-old woman with BLBC type breast cancer. The first four rows in Fig. 1 show the MRI, kinetic curve, and mammography images of these four patients. The last row shows the probability values of the masses margin on MRI, the kinetic curve in the delayed phase, and calculation distributions between the patients.

3.3. Decision tree and significant features

Of the 82 variables, 24 of the most important features (Importance = 1) were preliminarily chosen to build the decision tree. The names and importance of the variables are reported in the supplementary information later in this manuscript. Ultimately, the variables of the seven features were selected as nodes for classifying the corresponding molecular subtypes of breast cancer, consisting of the masses margin and calcification on mammography, and mass margin, kinetic curve in the delayed phase, mass internal enhancement characteristics, and non-mass enhancement distribution on MRI, and breastfeeding history. The distribution characteristics of these features are presented in Fig. 2. These characteristics were important for the decision-making process, as shown in Fig. 3 (adjusted *p*-values with Bonferroni corrections).

Calcifications: the HER2 subtype had the largest percentage of calcifications, whereas the BLBC subtype had the lowest. HER2 was associated with regional calcifications while Luminal A and B had more grouped calcifications.

Mass and non-mass margins: on mammography, the margins of the lesions of Luminal A and B tended to be microlobulated, whereas those of HER2 and BLBC tended to be indistinct. On MRI, Luminal A and B both exhibited irregular and spiculated patterns, but Luminal B had more instances of a spiculated pattern than Luminal A. BLBC showed more of a circumscribed pattern than the other molecular subtypes.

Mass and non-mass internal enhancement: the BLBC subtype showed circumscribed rim enhancement on MRI. HER2 showed more non-mass enhancement than the other subtypes.

Kinetic curve in the delayed phase on MRI: there were three patterns (i.e., persistent, plateau, and washout) of the kinetic parameters. BLBC showed more of the washout pattern, whereas HER2 had a more persistent kinetic curve in the delayed phase compared with the other molecular subtypes.

Breastfeeding history: breastfeeding history was an important factor in the differentiation between Luminal A and Luminal B on the decision tree since we observed less of a history of breastfeeding in Luminal B than in Luminal A.

Based on these features, rule sets were derived for Luminal A with a total of five rules; Luminal B had a total of three rules; HER2 had a total of three rules; BLBC had one rule. The rule sets are also reported in the supplementary information. Together, these rule sets formed the decision tree shown in Fig. 3. Similar to the way human radiologists interpret and make decisions, the decision tree can predict in detail the molecular subtypes of breast cancer using the probability of all the features.

3.4. Model evaluation

Compared with the pathological results, the total accuracy of the decision tree model was 74.1%. The confusion matrix between the predicted and actual values of the four molecular subtypes of all the samples (n = 363) is shown in Table 4. Average evaluation measures of the 10-fold cross-validation are shown in Table 5.

Through 10-fold cross-validation in mammography and MRI, the average accuracies of the training set in the decision tree model were 67.77% and 72.20%, respectively, while the average accuracies of the validation set were 67.71% and 70.31%, respectively. The accuracy of mammography was lower than that of MRI. The confusion matrix for each round is reported in the supplementary information. The average sensitivity, PPV, and F1-scores of each molecular subtype for the 10-fold cross-validation are shown in Table 6.

4. Discussion

To the best of our knowledge, this study is the first to report on using a combination of clinical history, radiologist-defined image feature extraction from both mammography and MRI, and decision tree-based machine learning to support the prediction of breast cancer molecular subtypes in a multi-modal setting. We developed and established a multi-parametric model based on fuzzy mathematics to solve the issue of cognitive differences between different radiologists and improve overall cognitive ability using swarm intelligence. In addition, the extracted features are standardized by following the BI-RADS standard to ensure that the study can be conveniently repeated and established in other medical settings.

Using a decision tree prediction model, we found the following diagnostically significant features: age, breastfeeding history, mass margin and calcification on mammography, and mass margin and internal enhancement, non-mass and internal enhancement distribution, and kinetic curve in the delayed phase on MRI. We found that the ages of onset of HER2 and BLBC subtypes were younger than the age of onset for the luminal subtypes ($p < 0.001$), which was not surprising because HER2 and BLBC were more malignant and progress more quickly [7]. Breastfeeding history differentiated between Luminal A and Luminal B on the decision tree ($p < 0.001$), a finding that was consistent with our previous study [25]. We found that calcifications were more frequent in the HER-2 subtype, which was concordant with a previous study [26]. Margins differentiated between the luminal cancers, which had irregular, spiculated margins with Luminal B having more spiculated margins than Luminal A, and HER2 and BLBC, which had indistinct margins. The findings regarding the luminal type margins agree with a study by Navarro Vilar L et al. [27], who also found that Luminal B margins were irregular and spiculated. However, our findings regarding HER2 margins were in contrast to Grimm LJ et al. [12] who found that HER2 cancers more frequently had a smooth margin than other subtypes. Enhancement differentiated between BLBC, which had rim enhancement, and HER2, which had non-mass enhancement, the latter of which was also found by Navarro Vilar L et al. [27]. Regarding kinetic curves, our results showed that the BLBC subtype showed more of a washout pattern whereas the HER2 subtype was more persistent in the delayed phase compared with other molecular subtypes. This finding was in contrast to Navarro Vilar L et al. [27], who found no statistically significant differences in the dynamic curves on MRI between different subtypes. Overall, these findings support the hypothesis that the molecular subtypes of breast cancer are often associated with particular imaging characteristics and that these characteristics can be used to differentiate between the molecular subtypes [10–15,26,27].

Through cross-validation, the average accuracy of the decision tree on MRI was higher than on mammography. This difference in accuracy is due to a well-known reason: MRI provides more information about soft tissue structure and pathology than mammography. Accuracy improved even more when information from both mammography and MRI were used together (see Table 6), which is not surprising since clinical decisions regarding breast cancer are currently made based on the results of both mammography and MRI, and not on one or the other.

The overall accuracy of our model was 74.1%, which was lower than that of existing radiomics studies, which were usually more than 90%. However, the high accuracies achieved by these other studies

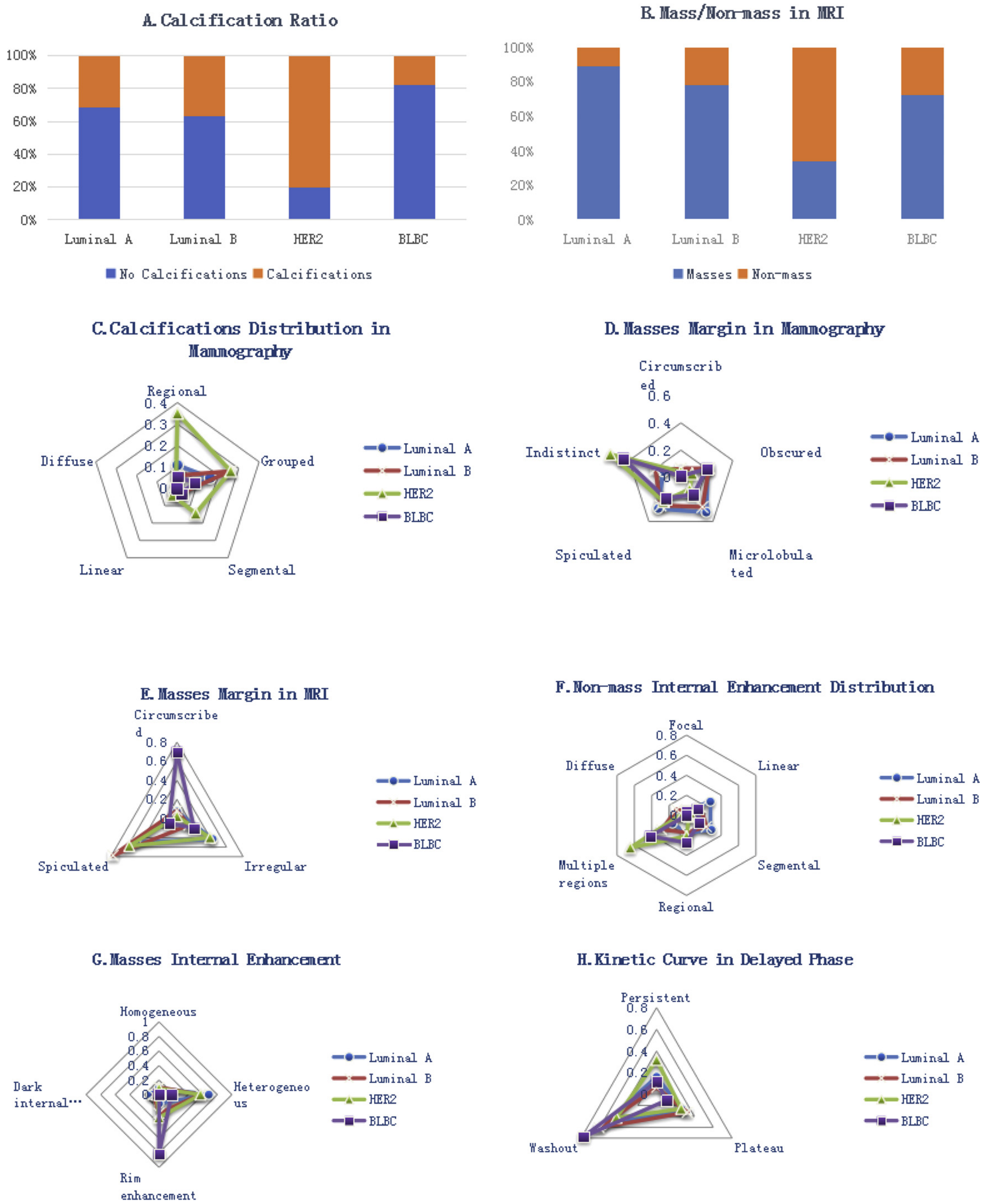

Fig. 2. The distribution characteristics of the eight most important features in the four molecular subtypes of breast cancer. A and B is the bar graph represents the variables with clear meaning without ambiguity. A shows the calcification ratio of lesions in the four molecular subtypes. B shows the non-mass/mass enhancement ratio on MRI in the four molecular subtypes. C ˜ H are the radar maps representing the average scores of fuzzy features by 5 radiologists. If the score of a molecular subtype tends to a certain angle of the radar map, it shows that it has a higher probability to correlate with this image feature. For example, C shows that the value of the HER2 subtype is more inclined to a regional calcifications distribution than other subtypes.

must be viewed with caution as these studies developed and tested their models on images with identical scanning parameters acquired on the same scanner. The stability of these models and consequently their corresponding accuracies were difficult to maintain on different scanners with different parameters [32]. For example, a review [32] of 41 papers on radiomics pointed out that the repeatability and reproducibility of radiomic features were inconsistently impacted by variable processing details, such as image acquisition settings, image reconstruction algorithms, digital image preprocessing, and software used to extract radiomic features. Our model was actually based on images from a variety of different scanners and combine BI-RADS features using supervised learning. For this reason, we believe that our model can produce comparable accuracies at different medical centers, allowing it to be clinically useful.

Our model was more accurate than that of Ha R et al. who proposed using a convolutional neural network (CNN), which is a deep learning method, to predict breast cancer molecular subtypes; this method only achieved an accuracy of 70% [30]. Most likely our superior accuracy is due to the fact that our study extracted image features that are clinically meaningful to radiologists, can be easily generalized, and are less

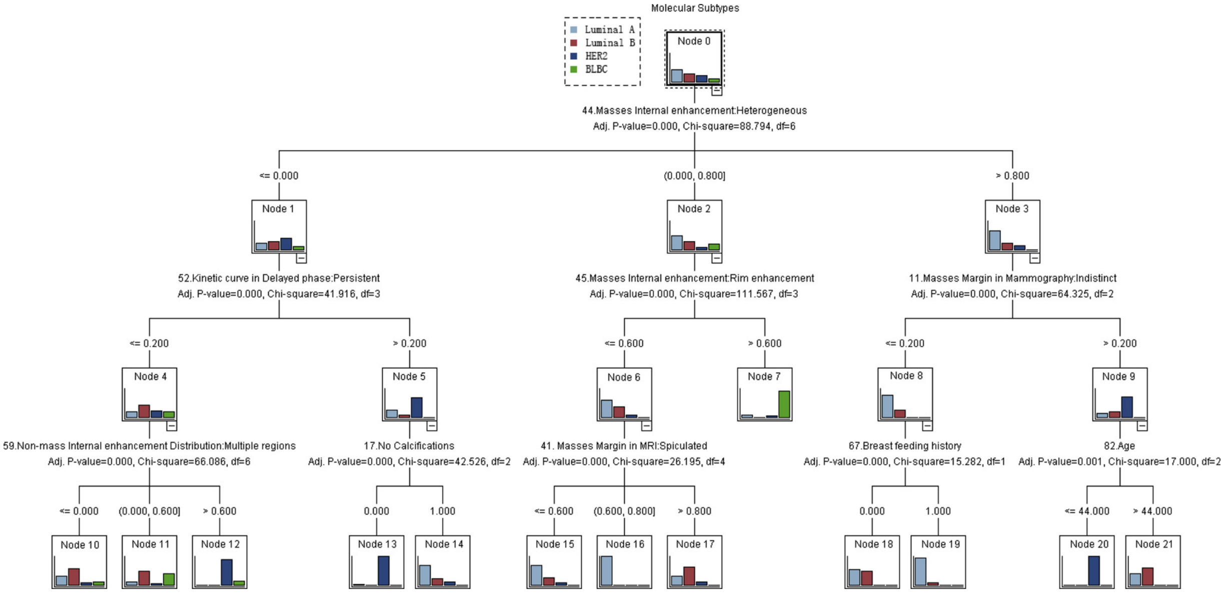
Molecular Subtypes
Luminal A
Luminal B
HER2
BLBC
Node 0
44.Masses Internal enhancement:Heterogeneous
Adj. P-value=0.000, Chi-square=88.794, df=6
<= 0.000
(0.000, 0.800]
> 0.800
Node 1
Node 2
Node 3
52.Kinetic curve in Delayed phase:Persistent
Adj. P-value=0.000, Chi-square=41.916, df=3
45.Masses Internal enhancement:Rim enhancement
Adj. P-value=0.000, Chi-square=111.567, df=3
11.Masses Margin in Mammography:Indistinct
Adj. P-value=0.000, Chi-square=64.325, df=2
<= 0.200
> 0.200
<= 0.600
> 0.600
<= 0.200
> 0.200
Node 4
Node 5
Node 6
Node 7
Node 8
Node 9
59.Non-mass Internal enhancement Distribution:Multiple regions
Adj. P-value=0.000, Chi-square=66.086, df=6
17.No Calcifications
Adj. P-value=0.000, Chi-square=42.526, df=2
41. Masses Margin in MRI:Spiculated
Adj. P-value=0.000, Chi-square=26.195, df=4
67.Breast feeding history
Adj. P-value=0.000, Chi-square=15.282, df=1
82.Age
Adj. P-value=0.001, Chi-square=17.000, df=2
<= 0.000
(0.000, 0.600]
> 0.600
0.000
1.000
<= 0.600
(0.600, 0.800]
> 0.800
0.000
1.000
<= 44.000
> 44.000
Node 10
Node 11
Node 12
Node 13
Node 14
Node 15
Node 16
Node 17
Node 18
Node 19
Node 20
Node 21

Fig. 3. The CHAID tree for the prediction of molecular subtypes in all 363 patients. The initial study samples (node 0, 151 Luminal A, 96 Luminal B, 76 HER2,40 BLBC) is split into child nodes (node 1–21) by the independent variable showing the highest discriminatory power based on chi-squared statistics with Bonferroni corrections. After 3 ramifications, the study sample is split into 12 terminal nodes (node 10–21) where no further differentiation can be achieved. In each node, bars indicate relative fractions of the 4 molecular subtypes with the black line on the left as the 100% denominator.

vulnerable to various image acquisition and post-processing methods while deep learning methods do not generate clinically meaningful image features for physicians to understand.

The main strengths of our study are that our model integrated numerous clinical and imaging features, that the imaging features are generalizable and can be recognized by radiologists across institutions worldwide, and that our model utilizes swarm intelligence. Our study extracted 82 standardized imaging features based on the BI-RADS criteria from both MRI and mammography and used decision tree classification to distinguish between different molecular subtypes of breast cancer whereas previous studies have only exploited a limited number of analytical MRI features on MRI, as evaluated by radiologists, and have only used decision tree classification methods to differentiate between benign and malignant breast lesions [28] and to predict lymph node metastases from breast cancer [31]. One study [28] argued that the BI-RADS criteria are of limited use and had not been shown to be more reliable than other diagnostic features. But we showed that the feature quantization method proposed in this study can improve the shortcomings of BI-RADS criteria.

Because all of the employed features are defined in the BI-RADS manual, they are easily generalizable and interpretable by radiologists using different scanners across multiple institutions whereas most previous studies based on radiomics features lack interpretability for radiologists and lack standardization in terms of applied features [18]. For example, the radiomics feature gray-level co-occurrence Matrix (GLCM) does not have a straightforward correlation to an image feature familiar to radiologists.

We believe image features identified by radiologists are highly stable regardless of different machines and imaging parameters because the human neural network can extract higher abstract features from images whereas machine learning algorithms, particularly the traditional methods, are sensitive to changes in imaging that the human brain can appropriately disregard. However, we understand that such a highly abstract extraction processes sacrifices mathematical precision and leads to inter-observer variability (Fig. 4). Previous studies address this variability by relying on consistency tests and third parties to broker a consensus between observers who disagree. However, we believe that the variability between observers reflects the nonlinearity and complexity of the real world thus should not be avoided, hence our use of swarm intelligence.

In summary, there are several differences between our study and previous studies that make our study unique. First, previous studies have different sample sizes and races while we focus on the Chinese population. Second and more importantly, previous studies [12,28] using BI-RADS features only performed correlation analyses; however, our study established a more complex prediction model using features extracted from mammography, MRI, and clinical metadata. Third, previous studies are vulnerable to radiologists' subjective errors when assessing imaging features whereas our method established a multiparametric model based on fuzzy mathematics to reduce cognitive differences.

Our study is limited by the fact that our decision tree model lacks advanced optimization. Our decision tree selected a limited number of nodes to establish a tree structure to avoid overfitting, which resulted in

Table 4
Confusion matrix between actual and predicted classification of molecular subtypes on all samples.

	Predicted-molecular subtypes				Total	Sensitivity	PPV	F1-score	Accuracy
Molecular subtypes	Luminal A	Luminal B	HER2	BLBC					
Luminal A	120	28	1	2	151	79.47%	75.47%	77.42%	
Luminal B	34	62	0	0	96	64.58%	55.86%	59.90%	
HER2	5	8	62	1	76	81.58%	95.38%	87.94%	
BLBC	0	13	2	25	40	62.50%	89.29%	73.53%	
Total	159	111	65	28	363				74.1%

Table 5
Average evaluation measures of 10-fold cross-validation on all samples.

	Training set				Validation set			
Average	Sensitivity	PPV	F1-score	Accuracy 78.70%	Sensitivity	PPV	F1-score	Accuracy 72.43%
Luminal A	82.81%	79.44%	80.49%		81.90%	74.82%	77.49%	
Luminal B	65.78%	71.02%	67.55%		58.86%	67.10%	60.96%	
HER2	79.72%	79.77%	78.83%		64.97%	72.07%	63.98%	
BLBC	54.60%	64.60%	57.41%		43.33%	55.67%	43.46%	

Table 6
Average evaluation measures of 10-fold cross-validation on mammography and MRI.

Mammography								
Train set					validation set			
Molecular Subtypes average	sensitivity	PPV	F1-score	Accuracy 67.77%	sensitivity	PPV	F1-score	Accuracy 67.71%
Luminal A	78.81%	73.46%	76.04%		79.45%	73.98%	75.88%	
Luminal B	43.75%	68.84%	53.48%		43.87%	67.74%	52.19%	
HER2	92.08%	59.31%	72.14%		90.46%	58.53%	70.34%	
BLBC	37.49%	68.24%	48.33%		36.67%	62.33%	42.52%	
MRI								
Train set					validation set			
Molecular Subtypes average	sensitivity	PPV	F1-score	Accuracy 72.20%	sensitivity	PPV	F1-score	Accuracy 70.31%
Luminal A	86.04%	69.56%	76.49%		82.85%	68.56%	73.33%	
Luminal B	58.73%	65.08%	60.21%		57.86%	62.42%	42.83%	
HER2	58.84%	94.68%	72.49%		57.72%	92.67%	69.43%	
BLBC	76.82%	72.86%	74.37%		78.83%	77.02%	61.09%	

the exclusion of some important collinearity features that are important factors in clinical practice. Thus, the structure of the decision tree model with various forms of representations should be further tested to select the most optimal features. In future studies, we plan to use a multi-test decision tree to further optimize the prediction model, taking full account of other low-ranked features to obtain more stable and reliable predictions.

5. Conclusion

We proposed a completely “white box” machine learning method to predict the molecular subtypes of breast cancers based on the BI-RADS feature description in a multi-modal setting. It is a promising approach that relies on discriminative feature selection and classification, which allows doctors to understand it easily so that they can integrate it into their clinical practice in order to make precise diagnoses and optimal therapeutic decisions.

Conflict of interest

None.

Acknowledgments

Grant Support: This work was supported by the Project of Science and Technology in Shenzhen, China (JCYJ20170307095935867 and

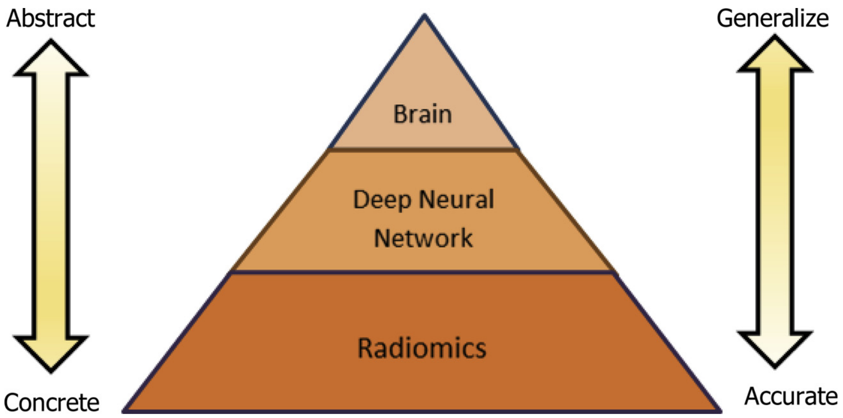

Fig. 4. Pyramid of Cognitive Ability. The human brain has the strongest abstraction and generalization ability, but its accuracy might be insufficient. Radiomics can accurately describe features, but it lacks the ability of abstraction and generalization. The feature extraction ability of the deep neural network lies between them.

JCYJ20180305164740612).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ejrad.2019.03.015.

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, CA Cancer J. Clin. 68 (2018) 7–30.
[2] A. Toss, M. Cristofanilli, Molecular characterization and targeted therapeutic approaches in breast cancer, Breast Cancer Res. 17 (2015) 60.
[3] C.M. Perou, T. Sørlie, M.B. Eisen, et al., Molecular portraits of human breast tumours, Nature 406 (2000) 747–752.
[4] K.E. Huber, L.A. Carey, D.E. Wazer, Breast cancer molecular subtypes in patients with locally advanced disease: impact on prognosis, patterns of recurrence, and response to therapy, Semin. Radiat. Oncol. 19 (2009) 204–210.
[5] A. Goldhirsch, W.C. Wood, A.S. Coates, et al., Strategies for subtypes–dealing with the diversity of breast cancer: highlights of the St. GAllen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011, Ann. Oncol. (22) (2011) 1736–1747.
[6] S. Park, J.S. Koo, M.S. Kim, et al., Characteristics and outcomes according to molecular subtypes of breast cancer as classified by a panel of four biomarkers using immunohistochemistry, Breast 21 (2012) 50–57.
[7] M.J. Engstrom, S. Opdahl, A.I. Hagen, et al., Molecular subtypes, histopathological grade and survival in a historic cohort of breast cancer patients, Breast Cancer Res. Treat. 140 (2013) 463–473.
[8] R. Rouzier, C.M. Perou, W.F. Symmans, et al., Breast cancer molecular subtypes respond differently to preoperative chemotherapy, Clin. Cancer Res. 11 (2005) 5678–5685.
[9] L.A. Carey, E.C. Dees, L. Sawyer, et al., The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes, Clin. Cancer Res. 13 (2007) 2329–2334.
[10] B.K. Killelea, A.B. Chagpar, J. Bishop, et al., Is there a correlation between breast cancer molecular subtype using receptors as surrogates and mammographic appearance? Ann. Surg. Oncol. 20 (2013) 3247–3253.
[11] A.G. Bitencourt, N.P. Pereira, L.K. França, et al., Role of MRI in the staging of breast cancer patients: does histological type and molecular subtype matter? Br. J. Radiol. 88 (2015) 20150458.
[12] L.J. Grimm, J. Zhang, J.A. Baker, et al., Relationships between MRI breast imaging-reporting and data system (BI-RADS) lexicon descriptors and breast cancer molecular subtypes: internal enhancement is associated with luminal B subtype, Breast J. 23 (2017) 579–582.
[13] M.S. Bae, M. Seo, K.G. Kim, et al., Quantitative MRI morphology of invasive breast cancer: correlation with immunohistochemical biomarkers and subtypes, Acta Radiol. 56 (2015) 269–275.
[14] E.J. Sutton, B.Z. Dashevsky, J.H. Oh, et al., Breast cancer molecular subtype classifier that incorporates MRI features, J. Magn. Reson. Imaging 44 (2016) 122–129.
[15] R. Ha, B. Jin, V. Mango, et al., Breast cancer molecular subtype as a predictor of the utility of preoperative MRI, AJR Am. J. Roentgenol. 204 (2015) 1354–1360.
[16] S. Yamamoto, D.D. Maki, R.L. Korn, et al., Radiogenomic analysis of breast cancer using MRI: a preliminary study to define the landscape, AJR Am. J. Roentgenol. 199 (2012) 654–663.
[17] M. Fan, H. Li, S. Wang, et al., Radiomic analysis reveals DCE-MRI features for prediction of molecular subtypes of breast cancer, PLoS One 12 (2017) e0171683.
[18] S.S. Yip, H.J. Aerts, Applications and limitations of radiomics, Phys. Med. Biol. 61 (2016) R150–66.
[19] B. Zhao, Y. Tan, W.Y. Tsai, et al., Reproducibility of radiomics for deciphering tumor phenotype with imaging, Sci. Rep. 6 (2016) 23428.
[20] N.J. Kleene, M.M. Michel, The capacity of trans-saccadic memory in visual search, Psychol. Rev. 125 (2018) 391–408.
[21] M. Czajkowski, M. Grzes, M. Kretowski, Multi-Test Decision Tree and Its Application to Microarray Data Classification vol. 61, (2014), pp. 35–44, https://doi.org/10.1016/j.artmed.2014.01.005.
[22] A. Goldhirsch, E.P. Winer, A.S. Coates, et al., Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013, Ann. Oncol. (24) (2013) 2206–2223.
[23] J. Krause, G.D. Ruxton, S. Krause, Swarm intelligence in animals and humans, Trends Ecol. Evol. 25 (2010) 28–34.
[24] P.R. Laughlin, E.C. Hatch, J.S. Silver, et al., Groups perform better than the best individuals on letters-to-numbers problems: effects of group size, J. Pers. Soc. Psychol. 90 (2006) 644–651.
[25] M. Wu, J. Ma, Association between imaging characteristics and different molecular subtypes of breast cancer, Acad. Radiol. 24 (2017) 426–434.
[26] B.K. Seo, E.D. Pisano, C.M. Kuzimak, et al., Correlation of HER-2/neu overexpression with mammography and age distribution in primary breast carcinomas, Acad. Radiol. 13 (2006) 1211–1218.
[27] L. Navarro Vilar, S.P. Alandete Germán, R. Medina García, et al., MR imaging findings in molecular subtypes of breast cancer According to BIRADS system, Breast J. 23 (2017) 421–428.
[28] P.A. Baltzer, M. Dietzel, T. Gröschel, et al., A simple and robust classification tree for differentiation between benign and malignant lesions in MR-mammography, Eur. Radiol. 23 (2013) 2051–2060.
[30] R. Ha, S. Mutasa, J. Karcich, et al., Predicting breast Cancer Molecular subtype with MRI dataset utilizing convolutional neural network algorithm, J. Digit. Imaging 31 (2019) [Epub ahead of print].
[31] M. Dietzel, P.A. Baltzer, T. Vag, et al., Application of breast MRI for prediction of lymph node metastases - systematic approach using 17 individual descriptors and a dedicated decision tree, Acta Radiol. 51 (2010) 885–894.
[32] A. Traverso, L. Wee, A. Dekker, R. Gillies, Repeatability and reproducibility of radiomic features: a systematic review, Int. J. Radiat. Oncol. Biol. Phys. 102 (2018) 1143–1158.