DOI: 10.4274/MMJ.galenos.2022.70094
Medeni Med J 2022;37:277-288

Original Article

Radiomics Features Based on MRI-ADC Maps of Patients with Breast Cancer: Relationship with Lesion Size, Features Stability, and Model Accuracy

Meme Kanserinde MRG-ADC Haritalarına Dayalı Radyomiks Özellikler: Lezyon Boyutu ile Stabilite ve Model Doğruluğunun İlişkisi

Begumhan BAYSAL[1], Hakan BAYSAL[2], Mehmet Bilgin ESER[1], Mahmut Bilal DOGAN[1], Orhan ALIMOGLU[2]

[1]Istanbul Goztepe Prof. Dr. Suleyman Yalcin City Hospital, Clinic of Radiology, Istanbul, Turkey
[2]Istanbul Goztepe Prof. Dr. Suleyman Yalcin City Hospital, Clinic of General Surgery, Istanbul, Turkey

ABSTRACT

Objective: To predict breast cancer molecular subtypes with neural networks based on magnetic resonance imaging apparent diffusion coefficient (ADC) radiomics and to detect the relation of lesion size with the stability of radiomics features.

Methods: This retrospective study included 221 consecutive patients (224 lesions) with breast cancer imaged between January 2015 and January 2020. Three sample size configurations were identified based on tumor size (experiment 1: all cases, experiment 2: >1 cm^3, and experiment 3: >2 cm^3). The tumors were segmented by three observers based on diffusion-weighted imaging-registered ADC maps, and the volumetric agreement of these segmentations was evaluated using the Dice coefficient. Stability of radiomics features (n=851) was evaluated with intraclass correlation coefficient (ICC, >0.75) and coefficient of variation (CoV, <0.15). Feature selection was made with variance inflation factor (VIF, <10) and least absolute shrinkage and selection operator regression. Outcomes were identified as molecular subtypes (Luminal A, Luminal B, HER2-enriched, triple-negative). Neural network performance was presented as an area under the curve and accuracies.

Results: Of the 851 radiomics features, 611 had ICC >0.75, and 37 remained stable in the first experiment, 49 in the second, and 59 in the third based on CoV and VIF analysis. High accuracy was demonstrated by the Luminal B, HER2-enriched, and triple-negative models in the first experiment (>80%), all models in the second experiment, and HER2-enriched and triple-negative models in the third experiment.

Conclusions: A positive stability is indicated by an increased lesion size related to radiomics features. Neural networks may predict moleculer subtypes of breast cancers over 1 cm^3 with high accuracy.

Keywords: Breast carcinoma, diffusion magnetic resonance imaging, computer-assisted image processing, machine learning, artificial intelligence

ÖZ

Amaç: Manyetik rezonans görüntülemede görünür difüzyon katsayısı (ADC) radyomiks verilerine dayalı sinir ağları ile meme kanseri moleküler alt tiplerini tahmin etmek ve lezyon boyutunun radyomiks özelliklerin stabilitesi ile ilişkisini saptamaktır.

Yöntemler: Bu retrospektif çalışma, Ocak 2015 ile Ocak 2020 tarihleri arasında görüntüleme yapılan meme kanserli hasta kohortunu (n=221, 224 lezyon) içermektedir. Lezyon boyutu ile radyomiks özelliklerinin stabilitesi arasındaki ilişkiyi incelemek için, tümör boyutuna dayalı (deney 1: tüm durumlar, deney 2: >1 cm^3 ve deney 3: >2 cm^3) üç grup oluşturulmuştur. Üç farklı gözlemci, difüzyon ağırlıklı görüntüleme ile oluşturulan ADC haritalarında tümörleri segmentlere ayırmış ve bu segmentasyonların hacimsel uyumu, Dice katsayısı kullanılarak değerlendirilmiştir. Radyomiks özellik (n=851) seçimi, sınıf içi korelasyon katsayısına (ICC), varyasyon katsayısına (CoV), varyans inflasyon faktörüne (VIF) ve en küçük mutlak küçülme ve seçim operatörü regresyonuna dayandırılmıştır. Sonuçlar moleküler alt tipler olarak tanımlanmıştır (Luminal A, Luminal B, HER2 ile zenginleştirilmiş ve üçlü negatif). Sinir ağı başarı performansı, eğri altındaki alan olarak sunulmuştur.

Bulgular: Sekiz yüz elli bir radyomiks özelliğinden 611'i ICC >0,75 idi. Bu özelliklerden CoV ve VIF analizi ile 37'si birincide, 49'u ikincide ve 59'u üçüncü deneyde sabit kaldı. İlk deneyde Luminal B, HER2 ile zenginleştirilmiş ve üçlü negatif alt tipler için geliştirilen tahmin modellerinin doğruluğu yüksekti (>%80). İkinci deneyde tüm modeller ve üçüncü deneyde ise HER2 ile zenginleştirilmiş ve üçlü negatif modeller yüksek doğruluğa sahipti.

Sonuçlar: Radyomiks özellikler, artan lezyon boyutuna bağlı olarak pozitif stabilite göstermektedir. Yapay sinir ağları, 1 cm^3 üzerindeki meme kanserlerini yüksek doğrulukla tahmin edebilmektedir.

Anahtar kelimeler: Meme kanseri, difüzyon manyetik rezonans görüntüleme, bilgisayar destekli görüntü işleme, makine öğrenimi, yapay zeka

Address for Correspondence: B. Baysal, Istanbul Goztepe Prof. Dr. Suleyman Yalcin City Hospital, Clinic of Radiology, Istanbul, Turkey
E-mail: baysalbegumhan@yahoo.com **ORCID ID:** orcid.org/0000-0003-0470-1683

Received: 19 July 2022
Accepted: 25 August 2022
Online First: 08 September 2022

Cite as: Baysal B, Baysal H, Eser MB, Dogan MB, Alimoglu O. Radiomics Features Based on MRI-ADC Maps of Patients with Breast Cancer: Relationship with Lesion Size, Features Stability, and Model Accuracy. Medeni Med J 2022;37:277-288

INTRODUCTION

Breast cancer is the most common cancer in women[1,2]. Remarkable developments in the fields of imaging, surgery, pathology, medical oncology, and genetics have led to a significant decline in breast cancer-related death rates in 30 years[1,3-5]. In current clinical practice, patients are classified based on their molecular subtypes[3,6], which can be identified via biopsy, thus informing treatment selection[6]. Currently, molecular subtypes guide treatment using tissue samples or by immunohistochemical markers[6,7]. However, the major challenges at this point are the limited volume of tumor represented in the biopsy sample and the reliability of the biopsy sample, especially in heterogeneous tumors[7]. A diagnostic prediction model may predict molecular subtypes using imaging data[8-10], for examples, automated artificial neural networks (ANN)[8] in which many networks can be trained with different configurations. In automation, the separation of the sample into training, test (hyperparameter tuning), and validation (hold-out) sets can be used to determine the training, error function, hidden activation, and output activation to be selected for the network structure[11], and human-induced bias is reduced[12]. However, explainability may be limited in nonlinear models.

Previous studies focused on radiomics features extracted from dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI)[10,13-19]. Although DCE-MRI is the most sensitive sequence in breast cancer imaging, the accumulation of gadolinium-based contrast agents in the brain has been a concern[20]. Additionally, using contrast agents in patients with obesity, diabetes, and renal failure requires great attention[1,20]. Therefore, non-contrast MRI protocols such as diffusion-weighted imaging (DWI) should be develope[5]. DWI delivers functional diffusivity data, and, with apparent diffusion coefficient (ADC) mapping, quantifies diffusivity related to cell density in solid tumors. DCE-MRI has been performed to predict molecular subtype, tumor histology, risk of recurrence, response to chemotherapy, and the probability of metastasis[8,10,13-17,19,21-23]. However, few studies used DWIs or ADCs radiomics as predictors[9,18,24].

Previous studies focusing on breast cancer molecular subtypes have not evaluated spatial overlap[10,13-19], and few have tested the interobserver reproducibility of the radiomics features[22,23]. Reproducibility is very important in radiomics feature extraction[25-27] and, along with sharing data, is as important as the studies' design, precision, and accuracy[26,28]. Moreover, to maintain the quality and reproducibility of the studies, studies on artificial intelligence that include complex models should report data with transparency.

The relation between the size of lesions and the stability of radiomics features has not been studied before. This study primarily aimed to predict breast cancer molecular subtypes with automated ANN created based on MRI ADC radiomics features and then to investigate the relationship between lesion size and stability of radiomics features and model accuracy.

MATERIALS and METHODS

Ethical Considerations

This retrospective study was approved by Local Ethics Committee of the Istanbul Medeniyet University Goztepe Training and Research Hospital (decision no: 2020/0303, date: 18.05.2020). The requirement for written informed patient consent was waived by the local ethics committee. We ensured adherence to the STARD 2015 statement[29] and the white papers and statements of European, United States, and Canadian societies[26,30-32]. The radiomics quality score was 18/36[33]. An Image Biomarker Standardization Initiative (IBSI)-compliant software was used for feature extraction[34].

Study Population and Data Collection

This model development study was conceived in Istanbul Medeniyet University Goztepe Prof. Dr. Suleyman Yalcin City Hospital. Data of patients histopathologically diagnosed with breast cancer in the general-surgery service of the university hospital between January 2015 and January 2020 were collected. The inclusion criteria were as follows: patients with breast MRI, including DWI sequences; detectable lesion on the DWI and ADC map; and invasive breast cancer as per the pathology report (if the patient had two different types of tumor histology, the tumors were included separately). The exclusion criteria were as follows: patients operated at the research center but without available imaging results and lesions detected on the ADC map but were pathologically diagnosed as *in situ* cancer. A complete pipeline is presented in Figure 1, and a flowchart of sample selection for the study is presented in E-Figure 1 (You can access all E-Figures and E-Tables from the link at the end of the article).

Based on the above criteria, 221 patients with 224 lesions were detected for the study (experiment 1, n=224). Breast cancer molecular subtypes based on pathological examination of surgery specimens were recorded in a worksheet by the surgery team. The analysis was repeated by narrowing the data set to over 1 cm^3 lesion (experiment 2, n=172) and over 2 cm^3 lesions (experiment 3, n=139). The 1.5 Tesla magnet power MRI protocols are described in E-Table 1.

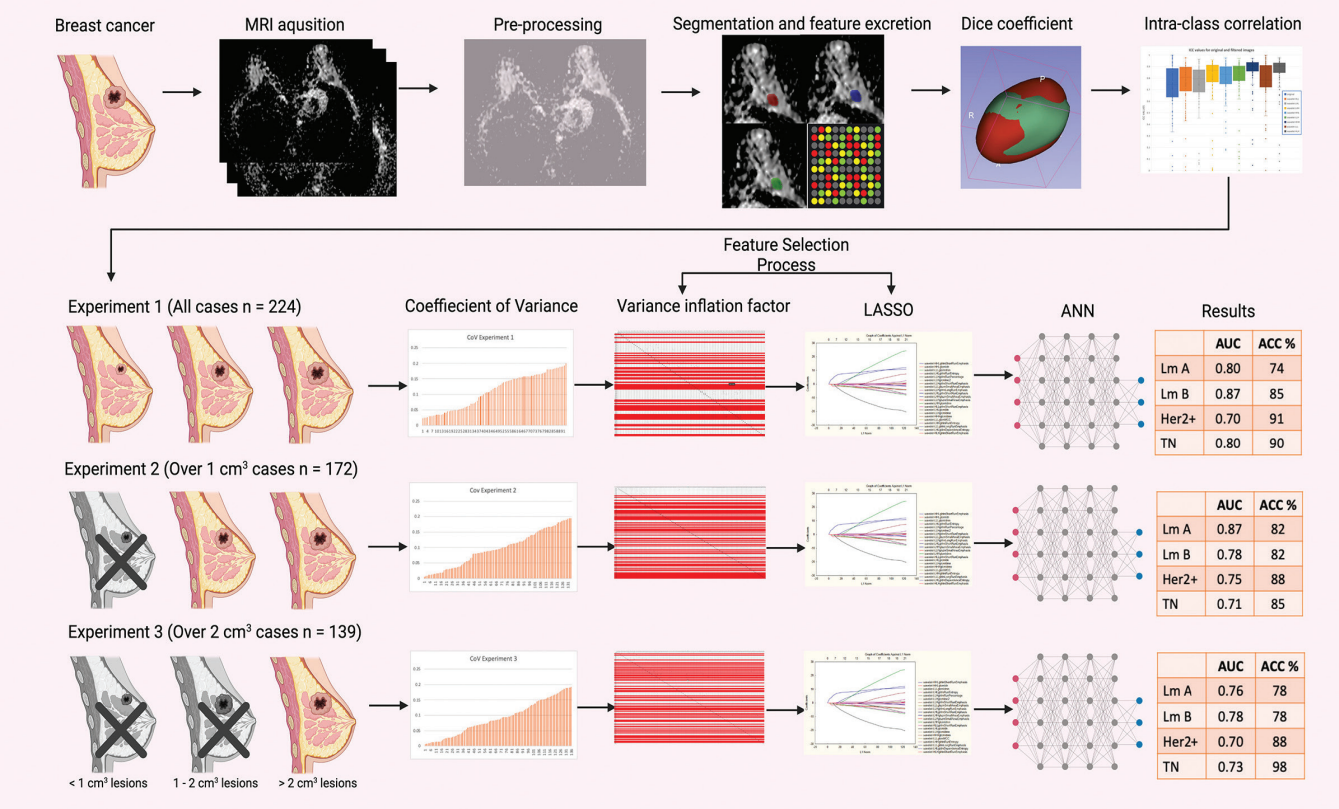

Figure 1. This scheme summarizes the entire study pipeline and results. Patients with MRI among the patients operated on for breast cancer in our hospital were included. After normalization and resampling, three observers independently segmented the lesions and obtained a radiomics feature. The agreements of the segmentations were tested with the Dice coefficient. Interobserver agreement for radiomics features was tested using intra-class correlation coefficient. Patient data were divided into three experimental groups based on the lesion size. Following the European Society of Radiology guideline, the pipeline coefficient variance and variance inflation factor analyses are also performed. As lesion size increased, lesion stability and the success of automated artificial neural networks also increase.

MRI: Magnetic resonance imaging, LASSO: Least absolute shrinkage and selection operator, ANN: Artificial neural networks, AUC: Area under the curve, ACC: Accuracy, HER2: Human epidermal growth factor receptor 2, TN: Triple-negative

Statistical Analysis

Predictors: Analysis of the ADC Maps

MRIs of the included patients were taken from the hospital archive and anonymized. Using the 3D Slicer (version 4.10.2; https://www.slicer.org) software, three radiologists with 8, 5, and 3 years of experience performed the segmentation from each axial segment where the tumor was located as seen on high b-value DWI images and verified on the ADC map[35,36]. Co-registration was made between T2 weighted images if the lesion was not detected from DWI. The predictor variables (radiomics features) of this study was extracted with PyRadiomics (version 2.2.0). All the radiomics features (n=851) were included, and wavelet-based filters were used. Raw ADC maps and resampled images (2.0x2.0x2.0 mm) were used and normalized[36,37]. Other detailed information about the radiomics features included in the study is provided in E-Table 2.

Outcomes

The outcomes of the study were molecular subtypes of breast cancer based on the biopsy: Luminal A, Luminal B, human epidermal growth factor receptor 2 (HER2)-enriched, and triple-negative (TN). These outcomes were coded as "one-vs-rest" orientation (E-Table 3)[27].

Features Stability Assessment

Interobserver agreement on the segmentations and radiomics features were evaluated by Dice similarity coefficient[27] and the intraclass correlation coefficient (ICC), respectively[38]. Discrepancies in the Dice similarity

Table 1. Clinicopathologic characteristics of the participants for three experiments.					
			All lesions (n=221, 224 lesions)	**Over 1 cm^3 (n=172, 172 lesions)**	**Over 2 cm^3 (n=139, 139 lesions)**
Age (mean ± SD, year)			54±11	55±12	54±12
Sex (female, n, %)			220, 99.6%	172, 100%	139, 100%
Median tumor size cm^3 (interquartile range)			1.57 (2.20)	3.37 (3.57)	5.56 (8.67)
Median voxel number (interquartile range)			167 (225)	347 (494)	650 (1264)
TNM classification	**T**	T1 (n, %)	79, 35%	56, 33%	43, 31%
		T2 (n, %)	113, 50%	97, 56%	79, 56%
		T3 (n, %)	16, 7%	11, 6%	11, 8%
		T4 (n, %)	6, 3%	5, 3%	4, 3%
	N	N0 (n, %)	113, 50%	87, 51%	73, 51%
		N1 (n, %)	64, 29%	52, 30%	39, 28%
		N2 (n, %)	21, 9%	17, 10%	15, 11%
		N3 (n, %)	18, 8%	15, 9%	12, 9%
	M	M0 (n, %)	212, 95%	168, 98%	136, 96%
		M1 (n, %)	3, 1%	2, 1%	2, 1%
Molecular subtypes					
Luminal A (n, %)			82, 37%	69, 40%	55, 39%
Luminal B (n, %)			76, 34%	55, 32%	45, 32%
HER2+ (n, %)			37, 17%	30, 17%	23, 16%
Triple negative (n, %)			23, 10%	18, 11%	13, 9%
SD: Standard deviation, HER2: Human epidermal growth factor receptor 2					

coefficient of segmentations (<0.50) were resolved by consensus. Features with an ICC >0.75 were further analyzed. Features from the three measurements were averaged, and the data were combined with the worksheet containing the pathology data. Worksheets containing three different sample size configurations were created. The coefficient of variation (CoV) analysis, which radiomics features showing >15% variance, was eliminated[26,28]. Then, Spearman's correlation (SC) analysis was performed to evaluate the remaining features, followed by variance inflation factor (VIF) analysis.

Collinearity-multicollinearity Analysis and Final Feature Selection

To achieve low collinearity-multicollinearity, VIF analyses were performed. In case VIF was above 10, the radiomic feature was eliminated[28,39]. For validation, SC analysis was performed between features and outcomes ($p<0.01$)[26]. Least absolute shrinkage and selection operator (LASSO) was used for future selection with random sampling method and 10-fold cross-validation.

Structuring Automated Artificial Neural Networks

Neural networks were binary classifiers as "one-vs-rest" orientation, and a diagnostic model was developed after feature selection using features selected for each experiment and outcome[27]. The data were divided into three subsamples in each training session using a random number generator. The software randomly sampled 50%-70% of the cases as training set, 10%-20% as a test set (hyperparameter tuning set), and 20%-30% as a validation (hold-out) set. Multilayer perceptron (MLP) and radial basis function (RBF) neurons were trained, and networks were feedforward and fully connected[11]. In each analysis, MLP or RBF neurons were trained, tested, and validated with unseen data set. The software automatically assigned the number of neurons (6-25), the number of layers [input layer (n= predictors for RBF and n= predictors + bias neuron for MLP), minimum two hidden layer, output layer (n=2, Positive event or not)], the number of bias neurons (minimum one per hidden layer both RBF and MLP), activation - hidden - output function [identity, logistic sigmoid, hyperbolic tangent,

Table 2. Stable radiomics features after coefficient of variance and variance inflation factor analyses for three experiments.

Experiment 1	Experiment 2	Experiment 3
1. wavelet - LLL GLRLM Short Run Emphasis, 2. wavelet - LLH GLRLM Short Run Emphasis, 3. wavelet - LLH GLRLM Run Percentage, 4. wavelet - LLL GLCM IMC2, 5. wavelet - LLH GLCM IMC2, 6. wavelet - LHL GLCM IDmn, 7. wavelet - HHH GLCM IDmn, 8. wavelet - HLH GLCM IDmn, 9. wavelet - LLL GLCM IDmn, 10. wavelet - HHL GLCM IDmn, 11. wavelet - LHH GLCM IDmn, 12. wavelet - LHH GLRLM Short Run Emphasis, 13. wavelet - LHL GLRLM Short Run Emphasis, 14. wavelet - LLL GLRLM Long Run Emphasis, 15. wavelet - LLH GLCM IDmn, 16. wavelet - LHL GLCM IDn, 17. wavelet - HHL GLCM IDn, 18. wavelet - HHH GLCM IDn, 19. wavelet - HLL GLRLM Short Run Emphasis, 20. wavelet - HLH GLRLM Short Run Emphasis, 21. wavelet - LLH GLSZM Small Area Emphasis, 22. wavelet - LLL GLSZM Small Area Emphasis, 23. wavelet - LLH GLRLM Long Run Emphasis, 24. wavelet - LLL GLCM MCC, 25. wavelet - HHH GLRLM Short Run Emphasis 26. wavelet - HHL GLRLM Short Run Emphasis, 27. wavelet - LLH GLCM MCC, 28. wavelet - LHL GLRLM Run Entropy, 29. wavelet - LHH GLRLM Run Entropy, 30. wavelet - LLL first order Entropy, 31. wavelet - LLL GLRLM Run Entropy, 32. wavelet - LLH GLRLM Run Entropy, 33. wavelet - LLH firstorder Entropy, 34. wavelet - LHH GLSZM Small Area Emphasis, 35. wavelet - HLH GLCM IMC2, 36. wavelet - LHL firstorder Entropy, 37. wavelet - LHL GLDM Dependence Entropy.	1. wavelet - HLL GLCM IDmn, 2. wavelet - LHL GLCM IDmn, 3. wavelet - LLL GLRLM Run Length Non-Uniformity Normalized, 4. wavelet - HHL GLCM IDmn, 5. wavelet - LLL GLCM IDmn, 6. wavelet - HHH GLCM IDmn, 7.wavelet - HLH GLCM IDmn, 8. wavelet - LHH GLCM IDmn, 9. wavelet - LLH GLCM IMC2, 10. wavelet - LHH GLRLM Short Run Emphasis, 11. wavelet - HHL GLCM IDn, 12. original GLCM IDn, 13. wavelet - LHL GLRLM Short Run Emphasis, 14. original GLRLM Run Length Non-Uniformity Normalized, 15. wavelet - LLH GLCM IDmn, 16. wavelet - HHH GLCM IDn, 17. wavelet - HLH GLRLM Short Run Emphasis, 18. wavelet - LLH GLSZM Small Area Emphasis, 19. original GLCM IMC, 20. wavelet - LHL GLDM Dependence Entropy, 21. wavelet - LHH GLDM Dependence Entropy 22. wavelet - HHL GLDM Dependence Entropy, 23.original GLSZM Small Area Emphasis, 24. wavelet - HLH GLDM Dependence Entropy, 25. wavelet - HLL GLDM Dependence Entropy, 26. wavelet - LLL GLSZM Zone Entropy, 27. original GLRLM Run Entropy, 28. wavelet - LLL GLCM MCC, 29. wavelet - LHL GLSZM Zone Entropy, 30. wavelet - LHH GLSZM Zone Entropy, 31. wavelet - HHH GLDM Dependence Entropy, 32. wavelet - LHL GLRLM Run Entropy, 33. wavelet - LLL GLCM Difference Entropy, 34. wavelet - HHL GLRLM Run Percentage, 35. wavelet - HLH GLSZM Zone Entropy, 36. wavelet - LLH GLCM MCC, 37. wavelet - HHL GLCM Inverse Variance, 38. wavelet - HLL GLSZM Zone Entropy, 39. wavelet - LHH GLSZM Small Area Emphasis, 40. original GLCM Difference Entropy, 41. wavelet - LHH GLRLM Run Entropy, 42. original GLCM Sum Entropy, 43. wavelet - LHL GLSZM Small Area Emphasis, 44. wavelet - HLL GLCM IMC, 45. wavelet - HLL GLRLM Run Entropy, 46. wavelet - LHL firstorder Entropy, 47. original shape Sphericity, 48. wavelet - LHH GLCM Joint Entropy, 49.wavelet - HLH GLCM IMC2.	1. wavelet - LLL GLRLM Short Run Emphasis, 2. wavelet - LLH GLRLM Short Run Emphasis, 3. wavelet - HLL GLCM IDmn, 4. wavelet - LHL GLCM IDmn, 5. wavelet - LLL GLCM IMC2, 6. original GLCM IDmn, 7. wavelet - HLH GLCM IDmn, 8. wavelet - LLL GLCM IDmn, 9. original GLRLM Short Run Emphasis, 10. wavelet - HHH GLCM IDmn, 11. wavelet - LHH GLCM IDmn, 12. wavelet - HHL GLCM IDmn, 13. wavelet - LLH GLCM IMC2, 14. wavelet - LLH GLCM IDmn, 15. wavelet - HHL GLCM IDn, 16. wavelet - LHH GLRLM Short Run Emphasis, 17. wavelet - HHH GLCM IDn, 18. wavelet - LHL GLRLM Short Run Emphasis, 19. wavelet - LLL GLRLM Long Run Emphasis, 20. wavelet - HLH GLRLM Short Run Emphasis, 21. wavelet - LLH GLSZM Small Area Emphasis, 22. original GLCM IMC2, 23. wavelet - LHL GLDM Dependence Entropy, 24. wavelet - LHH GLDM Dependence Entropy 26. wavelet - HLL GLDM Dependence Entropy, 27. wavelet - HHL GLDM Dependence Entropy, 28. wavelet - LLH GLRLM Run Entropy, 29. wavelet - HLH GLDM Dependence Entropy, 30. wavelet - LLL GLSZM Zone Entropy, 31. wavelet - LLL GLRLM Run Entropy, 32. wavelet - LLH GLRLM Long Run Emphasis, 33. wavelet - LHL GLSZM Zone Entropy, 34. original GLSZM Small Area Emphasis, 35. original GLRLM Run Entropy, 36. wavelet - LHH GLSZM Zone Entropy, 37. wavelet - HHH GLRLM Short Run Emphasis, 38. wavelet - LLH GLCM Sum Entropy, 39. wavelet - LLL GLCM Difference Entropy, 40. wavelet - HHH GLDM Dependence Entropy, 41. wavelet - LHL GLRLM Run Entropy, 42. wavelet - HLL GLSZM Zone Entropy, 43. wavelet - LLL GLCM MCC, 44. wavelet - HLH GLSZM Zone Entropy, 45. original GLRLM Long Run Emphasis, 46. wavelet - LLL GLCM Sum Entropy, 47. wavelet - HHL GLCM Inverse Variance, 48. original GLcm Difference Entropy, 49. wavelet - LHH GLRLM Run Entropy, 50. wavelet - LHH GLSZM Small Area Emphasis, 51. wavelet - LLH GLCM MCC, 52. wavelet - HLL GLRLM Run Entropy, 53. wavelet - LHL GLSZM Small Area Emphasis, 54. wavelet - LHH GLCM Joint Entropy, 55. wavelet - HLL GLCM IMC2, 56. wavelet - LLH GLCM Joint Entropy, 57. wavelet - HLH GLSZM Small Area Emphasis, 58. original shape Sphericity, 59. wavelet - HLL GLSZM Small Area Emphasis.

GLCM: Gray level co-occurrence matrix, GLRLM: Gray level run length matrix, GLSZM: Gray level size zone matrix, ID: Inverse difference, IMC: Informational measure of correlation, MCC: Maximal correlation coefficient

exponential, Softmax, and Gaussian (only available for RBF networks), and error function (sum of squares, cross entropy)], in these models by evaluating input, output data, and sub sample proportions[11]. Hyperparameter tuning was made with early-stopping algorithm[11]. Then, a neural network search was performed for each outcome in three sample size configurations. Figure 2 summarizes how automated ANN were trained, tested, and validated. Most accurate networks were retained for each experiment and outcome. Most efficient networks results are presented with area under the curve (AUC) (95% confidence intervals; lower and upper bounds)[26,27]. In receiver operating curve analysis, AUC >0.85 and p<0.01 is considered a validated classifier neural network. TIBCO Statistica version 13.5 (TIBCO Software, Palo Alto, CA) was used for statistical analyses and neural network training.

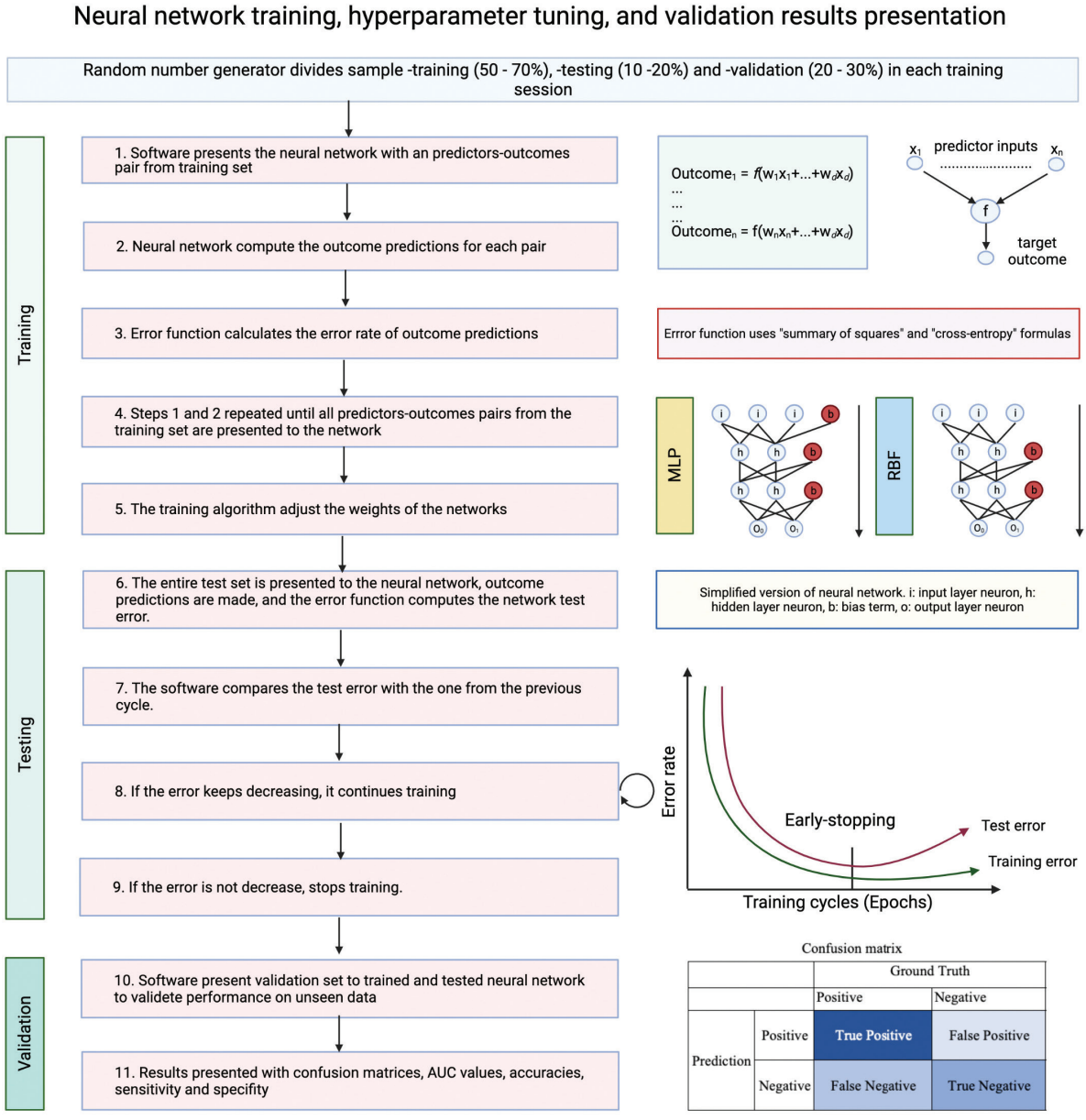

Figure 2. The diagram explains how automated artificial neural networks were trained, tested, and validated. Early-stopping hyperparameter tuning allows train neural networks faster, allowing training of thousands of neural networks in a short time.

AUC: Area under the curve, MLP: Multilayer perceptron, RBF: Radial basis function

RESULTS

Patient's Characteristics

This study included 221 patients (mean age, 54±11 years); of them, 220 (99%) were women. Clinicopathologic characteristics of the patients are presented in Table 1.

Feature Selection Results

The interobserver mean Dice coefficient values were as follows: between observers 1 and 2, 0.81±0.08 (0.80-0.82); for observers 1 and 3, 0.80±0.10 (0.79-0.82); and between observers 2 and 3, 0.73±0.11 (0.71-0.74).

The results of the resampled image features were not presented due to lower performance on ICC, CoV, VIF, LASSO analyses, and multivariate diagnostic models.

CoV, VIF, and LASSO regression analyses were performed separately in all three experiments (Figure 3). Of the 851 radiomic features, 611 were extracted from three segmentations with ICC values >0.75 and then included in a CoV analysis: 93 in the first experiment, 118 in the second experiment, and 136 in the third experiment. E-Table 3 presents an exact number of participants and outcome events for each analysis.

In the VIF analysis, other features were excluded from the models (Figure 4) and features showing collinearity-multicollinearity were excluded, resulting in 37 features in the first experiment, 49 in the second experiment, and 59 features in the third experiment (Table 2, E-Figure 2).

In the correlation analysis, for all SC 'r' for the first and second experiments, the radiomics features were not successful. In the third experiment, all SC 'r' were <0.40, and p<0.01 for 12 predictors for TN breast cancer.

LASSO regression was used for regularization, and the analysis results for each outcome are shared in Table 3.

Diagnostic Prediction Model Results

From the three experiments, each of 12 neural networks contained four multivariable binary classifier models (Table 4). Confusion matrix and detailed

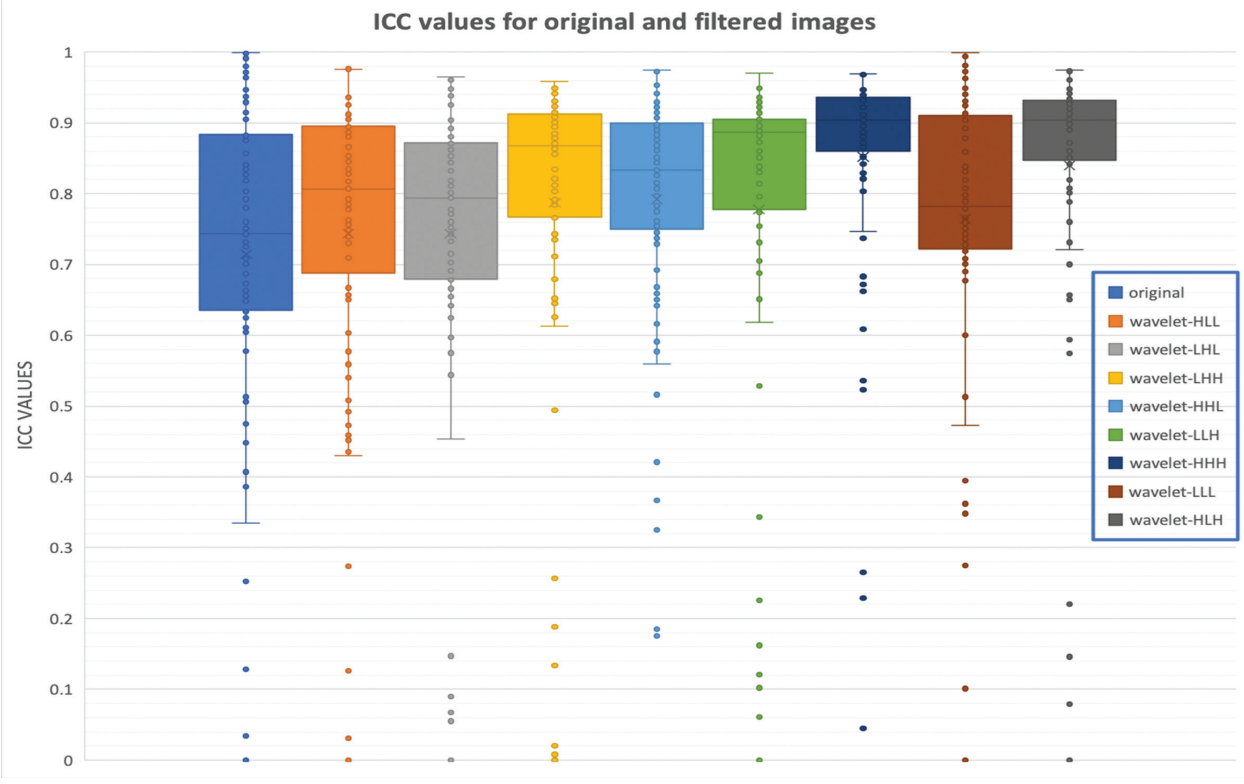

Figure 3. Evaluation of interobserver correlation coefficient (ICC) for original and filtered images. The first boxplot comes from the original image features, and only 46 features show high reproducibility (ICC >0.75). However, wavelet-HHH and wavelet-HLH features demonstrated higher reproducibility. Features from these filters have better reproducibility (80 and 79 features, respectively).

ICC: Interobserver correlation coefficient

performance metrics are presented for Luminal A in the E-Table 4, for Luminal B in the E-Table 5, for HER2-enriched in the E-Table 6, and for TN in the E-Table 7. In the validation (hold-out) set, the model trained for Luminal B in the first experiment and Luminal A in the second experiment reached AUC of 0.87 (0.73-0.99) and 0.87 (0.73-0.99), respectively. These findings indicate a high accuracy (>0.80) for Luminal B, HER2-enriched, and TN models in the first experiment; all models in the second experiment; and HER2-enriched, and TN models in the third experiment.

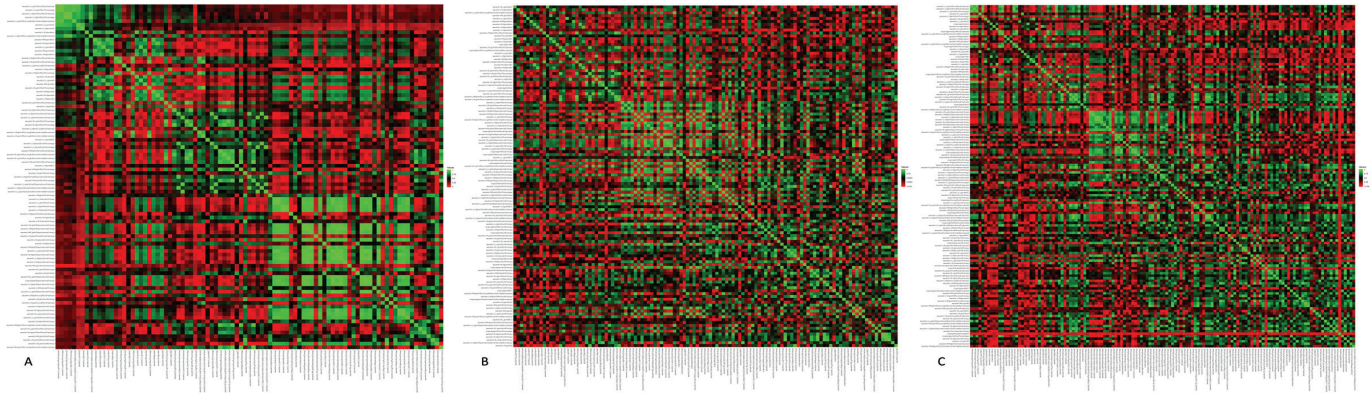

Figure 4. Heatmaps for the stable features before variance inflation factor analysis for three sample size configurations. The heatmap created for collinearity evaluation after features were eliminated due to interobserver agreement and high variance showed that there were the least stable features in the first configuration (A), and most of them were collinear. While the number of stable radiomics features increased in the second (B) and third (C) configurations, the collinearity decreased (heatmaps in detail for each experiment; https://github.com/MBE-hub/Breast/tree/master/Breast/4.Heat%20Maps).

Table 3. Selected predictor radiomics features analyzed by LASSO regression in three sample size configurations.

	Outcomes	Luminal A	Luminal B	HER2+ enriched	Triple-negative
Experiment 1	Intercept	9.967	-7.655	11.24	9.760
	Model λ	0.005	0.006	0.002	0.002
	% Deviation	0.082	0.091	0.085	0.139
	Predictors	4-6, 9-13, 15,16,18, 21-25, 28, 31, 34-36	3-6, 8-12, 15, 18, 19, 21-23, 25-27, 29, 31, 35, 36	1, 6, 8,9, 11, 12, 14-17, 21, 24-26, 28, 29, 33-37	3, 5, 7, 9, 11, 13-17, 19-24, 26, 28, 29, 34, 37
Experiment 2	Intercept	24.28	22.34	27.12	-45.10
	Model λ	0.008	0.007	0.005	0.005
	% Dev	0.117	0.160	0.183	0.186
	Predictors	1-4, 9, 12, 14-17, 19, 23, 25, 28, 29, 33, 34, 41, 43, 49	1, 4, 5, 8, 9, 12, 16, 17, 19, 23, 28, 29, 33, 34, 36, 38, 40, 43, 47, 49	1-3, 8, 9, 11-14, 21, 23, 26-29, 33, 36, 39, 42-44, 47	1, 7, 9, 10, 12, 14, 15, 27, 28, 33-35, 41-43, 47
Experiment 3	Intercept	-117.0	35.93	-18.59	-0.853
	Model λ	0.009	0.006	0.008	0.011
	% Dev	0.152	0.239	0.166	0.252
	Predictors	1, 5, 6, 12, 15, 17, 20, 23, 24, 30, 32-34, 36, 39, 44, 45, 48, 53, 54, 59	1, 4, 7, 11-13, 17, 21, 23, 26-28, 32, 34, 36, 39, 43, 45, 47, 58	1, 3-6, 9-11, 13, 15, 18, 19, 21, 34, 35, 43, 44, 47, 53, 58, 59	5, 6, 16, 17, 25, 26, 35, 37, 38, 47, 58

The most remarkable point about in the first configuration predictors is that all shape features are eliminated, and no feature without wavelet filters can pass the precision stage. Only eight features were extracted from the original image in the over 1 cm^3 and over 2 cm^3 configuration (Experiments 2 and 3). Only one of these is the shape feature (Sphericity). LASSO: Least absolute shrinkage and selection operator, HER2: Human epidermal growth factor receptor 2

Table 4. Artificial neural networks performance results of three experiments.

	Outcomes	Luminal A	Luminal B	HER2-enriched	Triple-negative
Experiment 1	Results				
	AUC*	0.71	0.79	0.76	0.78
	CI 95% lower	0.64	0.73	0.69	0.70
	CI 95% upper	0.78	0.85	0.84	0.87
	Acc (%)	64	77	84	90
	Sen (%)	10	50	38	26
	Spec (%)	96	91	93	97
	PPV (%)	62	75	54	55
	NPV (%)	64	77	88	92
Experiment 2	AUC*	0.65	0.87	0.86	0.90
	CI 95% lower	0.56	0.82	0.80	0.85
	CI 95% upper	0.73	0.92	0.92	0.95
	Acc (%)	65	84	86	94
	Sen (%)	13	75	57	56
	Spec (%)	100	89	92	98
	PPV (%)	100	76	61	77
	NPV (%)	63	88	91	95
Experiment 3	AUC*	0.86	0.77	0.82	0.88
	CI 95% lower	0.80	0.69	0.74	0.81
	CI 95% upper	0.92	0.85	0.89	0.95
	Acc (%)	83	76	87	96
	Sen (%)	73	67	65	69
	Spec (%)	90	81	91	98
	PPV (%)	83	64	60	82
	NPV (%)	83	83	93	97

HER2: Human epidermal growth factor receptor 2, AUC: Area under the curve, CI: Confidence interval, Acc: Accuracy, Sen: Sensitivity, Spec: Specificity, PPV: Positive predictive value, NPV: Negative predictive value. *All area under the curve p-values are <0.001. The first experiment had the largest sample size (n=221, 224 lesions), neurons trained with this data performed similarly to previous studies. Although the sample size was reduced, the high accuracy achieved in the second and third experiments may be associated with lesion size.

All data used in this study, the results of the analyses, and the trained neural network codes were shared publicly on GitHub (https://github.com/MBE-hub/Breast).

DISCUSSION

Based on the results of the present study, neural networks may predict molecular subtypes of breast cancer over 1 cm^3. Compared with previous studies, the present study evaluated the stability of radiomic features using the Dice similarity index, ICC, and CoV; the VIF was used to eliminate highly collinear features.

Currently, a minimally invasive approach is the most prevalent in medical practice[5]. Contrast-enhanced examinations are also considered an intervention[5,20]. Therefore, we focused on ADC radiomics as an alternative to invasive imaging modalities. Chen et al.[14] offered that ADC radiomics provided a more accurate diagnosis than DCE MRI radiomics. Unlike previous studies, the present study performed all breast cancers without limiting lesion size[9,10,13-16,22,23,35]. Experiment 2, which included over 1 cm^3, showed the best accuracy. In the first experiment, fewer pixels were segmented in small lesions, affecting the stability of the radiomics feature. Experiment 3 has a relatively limited sample size, and the validation set proportion was set to 30% to overcome this challenge and prevent overfitting.

Previous studies that assessed the molecular subtypes of breast cancer using an interobserver design did not evaluate the spatial overlap with the Dice

coefficient[9,10,13-16]. However, the evaluation of spatial overlap is recommended[27]. Traverso et al.[36,37] performed two studies based on cervix and rectal cancer using the Dice coefficient; the median Dice coefficient for the two observers was 0.73 and 0.75, respectively, which are similar to our results. The Dice coefficient provides a susceptible analysis as it depends on the pixel-to-pixel overlap of segmentation[36,37]. Given the sensitivity of the Dice method, the agreement of segmentations was almost perfect in this study. Furthermore, as mitigating certain discrepancies has been a challenge, observers re-evaluated the patients with a Dice coefficient <0.50 (n=9) to avoid bias.

In the present study, 72% of the radiomic features had ICC value >0.75. Using super-resolution ADC images, Fan et al.[22] performed a radiomics analysis to predict the histologic grade and Ki-67 expression status of breast cancer and found that shape and first-order features had an ICC >0.7, and neighborhood gray tone difference in matrix features showed large variance, with a low mean ICC. Zhang et al.[35] modelled multiparametric MRI to differentiate benign and malignant lesions from radiomics features and noted that all features had an ICC value of >0.75. Similarly, the present study included features with an ICC of >0.75. However, issues on reproducibility were raised due to not using an exclusion criterion for ICC[25,38] and not considering an interobserver assessment.

The European Society of Radiology (ESR) has recently published a statement on the validation of imaging biomarkers and described the validation pipeline[26]. The first step of this pipeline offers to evaluate features with a CoV analysis, stating, "high precision (low variance) is considered mandatory for the validation." Due to the novelty of this statement, none of the previous studies have used this analysis. In stability analysis, only 16% of features showed high stability even at the best condition (experiment 3: over 2 cm^3 lesions).

Parekh and Jacobs[19] reported that multivariable models had increased AUC (9-28%). Therefore, in the present study, we used multivariable models. Given the emerging use of multivariable regressions in feature selection tasks, collinearity-multicollinearity has become an essential problem. Kim[39] has described multicollinearity as a high degree of linear intercorrelation between predictor variables in a multivariable regression model. If collinearity is ignored, features on analysis become almost identical, thus increasing the relative error rate. In addition, features that better explain the model are ruled out due to the many identical features chosen. Although various methods have been defined, we preferred to eliminate features that show collinearity-multicollinearity in this study, and features were stable in only 7% of this elimination. Previous studies have not reported VIF analysis[9,10,13-16].

These results offered that radiomics features stability related to lesion size. The number of stable features increased with increasing lesion size. Despite the decrease in sample size in the second and third experiments, an increase in the Spearman correlation coefficient value, with an increase in the number of significant predictors, indicates a relationship between radiomics features stability and lesion size.

For validated biomarkers, the third item of the ESR statement pipeline requested that the p-value be <0.01 in the correlation analysis[26]. In the univariate analysis, a few radiomics features were validated in this study (E-Figure 2). Furthermore, their correlation coefficients were weak since LASSO regression was used for regularization in the current study[23-25].

Sutton et al.[10] have used the support vector machine, which included 38 features (mostly shape and contrast-enhancement patterns) and found the accuracy for prediction of TN molecular subtype breast cancer at 81%. This study used 851 features and found that all shape features, except for sphericity, are not stable in precision and accuracy. Moreover, sphericity could not be measured accurately, even in IBSI compliant software[34].

Diagnostic prediction models will benefit the clinician in detecting HER2-enriched and TN tumors. Leithner et al.[8] trained a TN ANN classifier with AUC =0.80 and 68.2% accuracy in the validation set. However, their other classifiers presented accuracies at 38.7%-70.3%. In the present study, neurons trained above 1 cm^3 configuration can estimate Luminal A, Luminal B, and HER2-enriched models with accuracy as high as that of the TN model; high specificity (>80%) was observed in the neurons in experiment 2, with moderate to high sensitivity (33%-80%).

The study has some limitations. The retrospective study design and single-center nature limit the generalizability of results. However, MR scans were performed with two different devices, and four different protocols and b-values in our center increased the potential diversity. For external validation, the cancer imaging archive was scanned, but without suitable data.

We used the manual segmentation method in this study because approximately 1/4 of our lesions were less than 1 cm^3. In addition, a recent study showed that automatic segmentation is not a good option for small lesions[40]. Fortunately, automated segmentation methods

have made rapid progress[21]. Future studies with large datasets may focus on breast cancer molecular subtype discrimination using convolutional neural networks and automated segmentation methods.

Using automated ANN minimizes human-induced bias[17]. However, the models created due to the weak linear relationship between predictors and outcomes reduce the network explainability. In the preliminary stage of the study, we also experienced machine learning methods such as support vector machines and K-nearest neighbors, and we attempted to train multiclass classifier methods such as gradient descent boosting and adaptive boosting. However, all these machine learning algorithms showed obviously lower accuracy than the models used in this study.

Radiomics features extraction yields the best results on iso-voxel partitioned images. Therefore, this study used both raw data (highly interpolated) and 2.0 mm iso-voxel images. Contrary to expectations, raw images showed better performance in this study, which was not supported by the literature, thereby limiting our discussion[8-10]. Based on our findings, high interpolation and high slice thickness caused artificial homogeneity on the resampled images, making the model success not better than raw images. Future studies should aim to increase the stability radiomics features and model success. Especially for DWI and ADC, it is necessary to increase the matrix values, decrease the section thickness, and increase the signal-to-noise ratio.

CONCLUSION

The stability of radiomics features is positively correlated to an increased lesion size. A diagnostic prediction model is a triaging and expediting the need for biopsy and/or for supporting histopathologic results in equivocal cases. However, while this prediction does not replace biopsy, it may require the triage of patients to be prioritized in radiology reporting, biopsy, and pathology reporting. The rapid and accurate triage of breast cancer molecular subtypes using imaging will be a potential development.

Acknowledgements: The authors thank Prof. Handan Ankaralı from Istanbul Medeniyet University Faculty of Medicine, Division of Biostatistics and Medical Informatics for her consultancy during the data generation, evaluation, and validation process.

Ethics

Ethics Committee Approval: This retrospective study was approved by Local Ethics Committee of the Istanbul Medeniyet University Goztepe Training and Research Hospital (decision no: 2020/0303, date: 18.05.2020).

Informed Consent: The requirement for written informed patient consent was waived by the local ethics committee.

Peer-review: Externally and internally peer-reviewed.

Author Contributions

Surgical and Medical Practices: H.B., O.A., Concept: B.B., H.B., M.B.E., M.B.D., O.A., Design: B.B., H.B., M.B.D., Data Collection and/or Processing: B.B., M.B.E., M.B.D., Analysis and/or Interpretation: B.B., M.B.E., O.A., Literature Search: B.B., H.B., M.B.E., M.B.D., O.A., Writing: B.B., H.B., M.B.E., M.B.D., O.A.

Conflict of Interest: The authors have no conflict of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES

1. DeSantis CE, Ma J, Gaudet MM, et al. Breast cancer statistics, 2019. CA Cancer J Clin. 2019;69:438-51.
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. Jin YH, Hua QF, Zheng JJ, et al. Diagnostic Value of ER, PR, FR and HER-2-Targeted Molecular Probes for Magnetic Resonance Imaging in Patients with Breast Cancer. Cell Physiol Biochem. 2018;49:271-81.
4. Lee SH, Park H, Ko ES. Radiomics in breast imaging from techniques to clinical applications: A review. Korean J Radiol. 2020;21:779-92.
5. Telegrafo M, Rella L, Stabile Ianora AA, Angelelli G, Moschetta M. Unenhanced breast MRI (STIR, T2-weighted TSE, DWIBS): An accurate and alternative strategy for detecting and differentiating breast lesions. Magn Reson Imaging. 2015;33:951-5.
6. Goldhirsch A, Winer EP, Coates AS, 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. 2013;24:2206-23.
7. Pisco AO, Huang S. Non-genetic cancer cell plasticity and therapy-induced stemness in tumor relapse: 'What does not kill me strengthens me'. Br J Cancer. 2015;112:1725-32.
8. Leithner D, Mayerhoefer ME, Martinez DF, et al. Non-Invasive Assessment of Breast Cancer Molecular Subtypes with Multiparametric Magnetic Resonance Imaging Radiomics. J Clin Med. 2020;9:1853.
9. Leithner D, Bernard-Davila B, Martinez DF, et al. Radiomic Signatures Derived from Diffusion-Weighted Imaging for the Assessment of Breast Cancer Receptor Status and Molecular Subtypes. Mol Imaging Biol. 2020;22:453-61.

10. Sutton EJ, Dashevsky BZ, Oh JH, et al. Breast cancer molecular subtype classifier that incorporates MRI features. J Magn Reson Imaging. 2016;44:122-9.
11. Statistica Automated Neural Networks (SANN) - Neural Networks Overview. Available from: https://docs.tibco.com/data-science/GUID-F60C241F-CD88-4714-A8C8-1F28473C52EE.html Accessed 19 Sep 2021
12. Sengupta E, Garg D, Choudhury T, Aggarwal A. Techniques to elimenate human bias in machine learning. In: Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends (SMART). IEEE; 2018. p. 226-30.
13. Chang RF, Chen HH, Chang YC, Huang CS, Chen JH, Lo CM. Quantification of breast tumor heterogeneity for ER status, HER2 status, and TN molecular subtype evaluation on DCE-MRI. Magn Reson Imaging. 2016;34:809-19.
14. Chen X, Chen X, Yang J, Li Y, Fan W, Yang Z. Combining Dynamic Contrast-Enhanced Magnetic Resonance Imaging and Apparent Diffusion Coefficient Maps for a Radiomics Nomogram to Predict Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer Patients. J Comput Assist Tomogr. 2020;44:275-83.
15. Grimm LJ, Zhang J, Mazurowski MA. Computational approach to radiogenomics of breast cancer: Luminal A and luminal B molecular subtypes are associated with imaging features on routine breast MRI extracted using computer vision algorithms. J Magn Reson Imaging. 2015;42:902-7.
16. Li H, Zhu Y, Burnside ES, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. NPJ Breast Cancer. 2016;2:16012.
17. O'Flynn EA, Collins D, D'Arcy J, Schmidt M, de Souza NM. Multiparametric MRI in the early prediction of response to neoadjuvant chemotherapy in breast cancer: Value of non-modelled parameters. Eur J Radiol. 2016;85:837-42.
18. Parekh VS, Jacobs MA. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. NPJ Breast Cancer. 2017;3:43.
19. Parekh VS, Jacobs MA. Multiparametric radiomics methods for breast cancer tissue characterization using radiological imaging. Breast Cancer Res Treat. 2020;180:407-21.
20. Gulani V, Calamante F, Shellock FG, Kanal E, Reeder SB; International Society for Magnetic Resonance in Medicine. Gadolinium deposition in the brain: summary of evidence and recommendations. Lancet Neurol. 2017;16:564-70.
21. Bhattacharjee R, Douglas L, Drukker K, Hu Q, Fuhrman J, Sheth D, Giger M. Comparison of 2D and 3D U-Net breast lesion segmentations on DCE-MRI. In Medical Imaging 2021: Computer-Aided Diagnosis. SPIE; 2021. p. 81-7.
22. Fan M, Yuan W, Zhao W, et al. Joint Prediction of Breast Cancer Histological Grade and Ki-67 Expression Level Based on DCE-MRI and DWI Radiomics. IEEE J Biomed Health Inform. 2020;24:1632-42.
23. Zhang Q, Peng Y, Liu W, et al. Radiomics Based on Multimodal MRI for the Differential Diagnosis of Benign and Malignant Breast Lesions. J Magn Reson Imaging. 2020;52:596-607.
24. Bickelhaupt S, Paech D, Kickingereder P, et al. Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography. J Magn Reson Imaging. 2017;46:604-16.
25. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: a practical guide for beginners. Diagn Interv Radiol. 2019;25:485-95.
26. European Society of Radiology (ESR). ESR Statement on the Validation of Imaging Biomarkers. Insights Imaging. 2020;11:76.
27. Erickson BJ, Kitamura F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. Radiol Artif Intell. 2021;3:e200126.
28. Kocak B, Kus EA, Kilickesmez O. How to read and review papers on machine learning and artificial intelligence in radiology: a survival guide to key methodological concepts. Eur Radiol. 2021;31:1819-30.
29. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. Radiology. 2015;277:826-32.
30. Jaremko JL, Azar M, Bromwich R, et al. Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. Can Assoc Radiol J. 2019;70:107-18.
31. Geis JR, Brady AP, Wu CC, et al. Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. Radiology. 2019;293:436-40.
32. European Society of Radiology (ESR). What the radiologist should know about artificial intelligence - an ESR white paper. Insights Imaging. 2019;10:44.
33. Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol. 2020;30:523-36.
34. Fornacon-Wood I, Mistry H, Ackermann CJ, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. Eur Radiol. 2020;30:6241-50.
35. Zhang Y, Zhu Y, Zhang K, et al. Invasive ductal breast cancer: preoperative predict Ki-67 index based on radiomics of ADC maps. Radiol Med. 2020;125:109-16.
36. Traverso A, Kazmierski M, Welch ML, et al. Sensitivity of radiomic features to inter-observer variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients. Radiother Oncol. 2020;143:88-94.
37. Traverso A, Kazmierski M, Shi Z, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. Phys Med. 2019;61:44-51.
38. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15:155-63. Erratum in: J Chiropr Med. 2017;16:346.
39. Kim JH. Multicollinearity and misleading statistical results. Korean J Anesthesiol. 2019;72:558-69.
40. Vorontsov E, Cerny M, Régnier P, et al. Deep Learning for Automated Segmentation of Liver Lesions at CT in Patients with Colorectal Cancer Liver Metastases. Radiol Artif Intell. 2019;1:180014.

E-Figures and E-Tables Link: http://glns.co/3102h