Available at www.sciencedirect.com

ScienceDirect

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

Original Research Article

Applications of machine-learning algorithms for prediction of benign and malignant breast lesions using ultrasound radiomics signatures: A multi-center study

Hassan Homayoun [a], *Wai Yee Chan* [k], *Taha Yusuf Kuzan* [c], *Wai Ling Leong* [b], *Kübra Murzoglu Altintoprak* [c], *Afshin Mohammadi* [d], *Anushya Vijayananthan* [b], *Kartini Rahmat* [b], *Sook Sam Leong* [e], *Mohammad Mirza-Aghazadeh-Attari* [f], *Sajjad Ejtehadifar* [d], *Fariborz Faeghi* [g], *U. Rajendra Acharya* [h,i,j], *Ali Abbasian Ardakani* [g,*,1]

[a] *Urology Research Center, Tehran University of Medical Sciences, Tehran, Iran*
[b] *Department of Biomedical Imaging, Universiti Malaya Research Imaging Centre, Faculty of Medicine, Universiti Malaya, Malaysia*
[c] *Department of Radiology, Sancaktepe Sehit Prof. Dr. Ilhan Varank Training and Research Hospital, Istanbul, Turkey*
[d] *Department of Radiology, Faculty of Medicine, Urmia University of Medical Science, Urmia, Iran*
[e] *Centre of Medical Imaging, Faculty of Health Sciences, Universiti Teknologi MARA Selangor, Bandar Puncak Alam, Selangor, Malaysia*
[f] *Medical Radiation Sciences Research Group, Faculty of Medicine, Tabriz University of Medical Sciences, Tabriz, Iran*
[g] *Department of Radiology Technology, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[h] *Ngee Ann Polytechnic, Department of Electronics and Computer Engineering, Singapore*
[i] *Department of Biomedical Engineering, School of Science and Technology, SUSS University, Singapore*
[j] *Department of Biomedical Informatics and Medical Engineering, Asia University, Taichung, Taiwan*
[k] *Gleneagles Hospital Kuala Lumpur, Department of Radiology, Jln Ampang, Kampung Berembang, Kuala Lumpur, Malaysia*

ARTICLE INFO

Article history:
Received 8 February 2022
Received in revised form
5 July 2022
Accepted 14 July 2022
Available online 22 July 2022

ABSTRACT

Artificial intelligence (AI) algorithms have an enormous potential to impact the field of radiology and diagnostic imaging, especially the field of cancer imaging. There have been efforts to use AI models to differentiate between benign and malignant breast lesions. However, most studies have been single-center studies without external validation. The present study examines the diagnostic efficacy of machine-learning algorithms in differentiating benign and malignant breast lesions using ultrasound images. Ultrasound images of 1259 solid non-cystic lesions from 3 different centers in 3 countries (Malaysia, Turkey, and Iran) were used for the machine-learning study. A total of 242 radiomics features were

[*] Corresponding author at: Department of Radiology Technology, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
E-mail addresses: A.ardekani@live.com, Ardakani@sbmu.ac.ir (A.A. Ardakani).

[1] ORCID: 0000-0001-7536-0973.

https://doi.org/10.1016/j.bbe.2022.07.004

Keywords:
Artificial Intelligence
Breast Cancer
Computer-assisted Diagnosis
Machine Learning
Radiomics

extracted from each breast lesion, and the robust features were considered for models' development. Three machine-learning algorithms were used to carry out the classification task, namely, gradient boosting (XGBoost), random forest, and support vector machine. Sensitivity, specificity, accuracy, and area under the ROC curve (AUC) were determined to evaluate the models. Thirty-three robust features differed significantly between the two groups from all of the features. XGBoost, based on these robust features, showed the most favorable profile for all cohorts, as it achieved a sensitivity of 90.3%, specificity of 86.7%, the accuracy of 88.4%, and AUC of 0.890. The present study results show that incorporating selected robust radiomics features into well-curated machine-learning algorithms can generate high sensitivity, specificity, and accuracy in differentiating benign and malignant breast lesions. Furthermore, our results show that this optimal performance is preserved even in external validation datasets.

1. Introduction

Computer-assisted diagnostic algorithms have the potential to significantly impact the workflow of radiology departments and the overall value chain of medical imaging [1,2]. Currently, the most meaningful additions to everyday clinical practice are being made in interpreting medical images and the immediately related tasks, such as triage of patients and optimization of imaging protocols [3]. Based on available published data and surveys of radiologists, breast imaging is the frontrunner of the race for artificial intelligence (AI) based automation in radiology [4], as a significant burden of tasks in breast imaging revolves around differentiating benign and malignant lesions in a background composed of relatively simple anatomy. AI algorithms have shown promise in answering yes or no diagnostic questions [5]. Until now, most commercially available AI solutions have been designed to decipher mammography images, an extension of X-ray imaging [6], and other imaging modalities have been partially neglected [7]. However, the trend is changing as more clinical studies aim to use chest CT and breast MRI, ultrasound, and PET imaging as inputs for AI algorithms. These algorithms not only diagnose breast lesions but may also give prognostications on disease outcome and risk stratification of malignant pathologies and be used in mass screening schemes of cardiothoracic conditions or follow-up of chronic diseases [8–10]. Noteworthy, AI algorithms with ultrasound image inputs are of significant importance, as breast ultrasound is the primary modality of imaging in many clinical scenarios [11] and is also used for unconventional means in many resources with scarce settings [12]. Ultrasound examination has been shown to be cost-efficient in many different settings, and unlike modalities such as MRI or PET imaging is indicated in a wide group of patients with suspected breast malignancies [13,14]. Furthermore breast ultrasound is enjoying a growing prominence in the study of dense breast tissue in addition to mammography, which may prove vital in correct classification of malignant lesions in such breast [15–17], and is also being further considered as a standalone modality in breast cancer imaging [18]. All of the mentioned reasons make ultrasound imaging a lucrative subject for quantitative imaging and computer assisted diagnosis systems [19].

However, certain obstacles limit AI clinical applicability to ultrasound images for breast lesion classification. One of the most important obstacles is the lack of studies that have utilized algorithms trained and tested on multiple datasets and validated using external data. External validation makes sure that the AI model is generalizable and can be used in real word diagnostic practices. Furthermore, the inclusion of data from different centers enables the model to train and ultimately be tested on a broad spectrum of pathologic lesions, increasing the diagnostic accuracy of the model [20].

Another issue is the lack of standardized protocols for developing and testing these algorithms. Recently, some initiatives have been to issue protocols, guidelines, and checklists to create specific AI models that rely on radiomics features [21]. These quantitative features are extracted from medical images and are used alongside AI models as proxy indicators of disease progression, chemo-resistance, disease survival, and more [22]. The present study aims to 1) introduce a robust set of radiomics features extracted from ultrasound images, which can be used in future multi-center studies, and 2) investigate the diagnostic profile of AI models based on radiomics features on multi-national datasets.

2. Patients and methods

2.1. *Patients and ultrasound imaging*

Ultrasound images of 1259 solid non-cystic lesions from 3 different centers in 3 countries were used for the AI study: 1) Malaysia, containing 853 lesions (501 benign and 352 malignant); 2) Iran, containing 232 lesions (109 benign and 123 malignant); and 3) Turkey, containing 174 lesions (99 benign and 75 malignant). Fig. 1 presents more information on patient selection, exclusion criteria of the study, and lesions' distributions among the centers.

For each center, only lesions with a definite diagnosis proven by histopathology (either needle, core, or open biopsies) were included. Benign lesions are determined as those that belong to the ICD-10-CM D24.1 (version 2022), and malignant

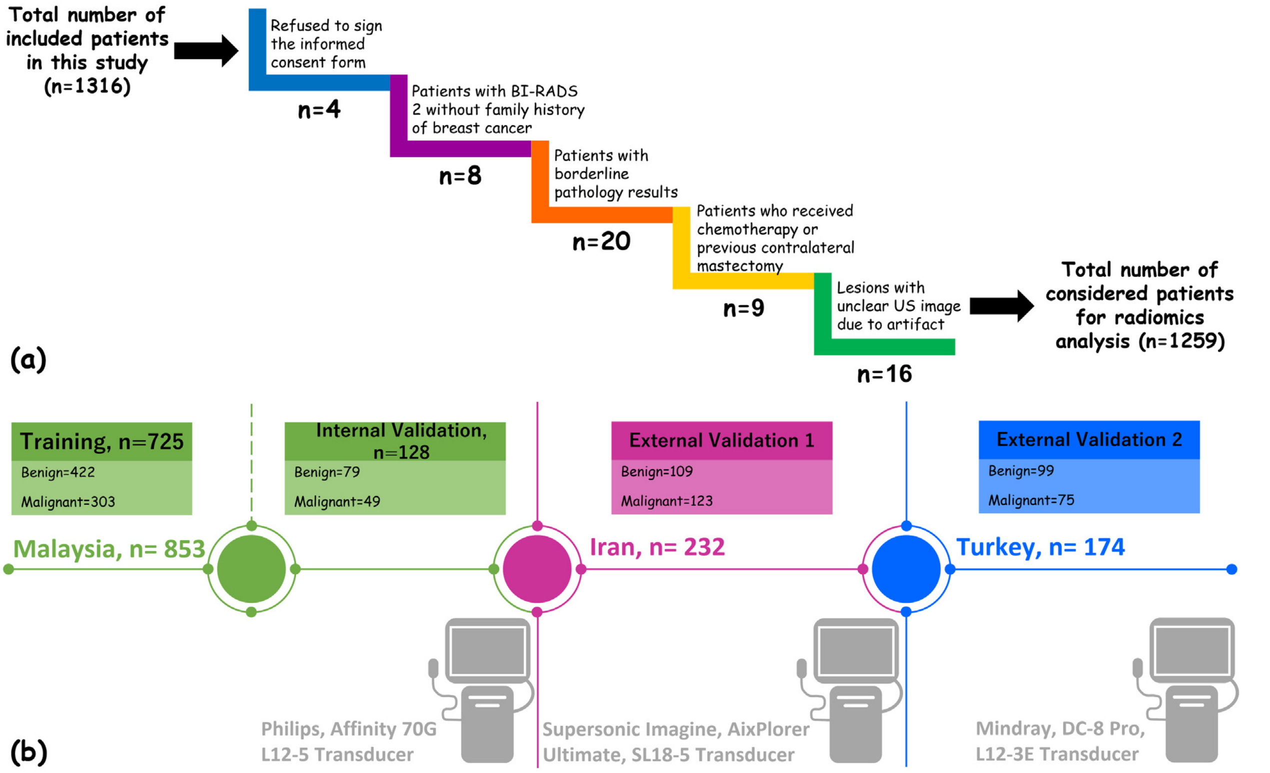

Total number of included patients in this study (n=1316)
Refused to sign the informed consent form
n=4
Patients with BI-RADS 2 without family history of breast cancer
n=8
Patients with borderline pathology results
n=20
Patients who received chemotherapy or previous contralateral mastectomy
n=9
Lesions with unclear US image due to artifact
n=16
Total number of considered patients for radiomics analysis (n=1259)
(a)
Training, n=725
Benign=422
Malignant=303
Internal Validation, n=128
Benign=79
Malignant=49
External Validation 1
Benign=109
Malignant=123
External Validation 2
Benign=99
Malignant=75
Malaysia, n= 853
Iran, n= 232
Turkey, n= 174
Philips, Affinity 70G L12-5 Transducer
Supersonic Imagine, AixPlorer Ultimate, SL18-5 Transducer
Mindray, DC-8 Pro, L12-3E Transducer
(b)

Fig. 1 – Exclusion criteria and patients selection process at a glance (a) and distribution of patients and the corresponding ultrasound machines used for each center in this study (b).

lesions are defined as lesions classified under the ICD-10 C50 code (version 2019, latest available update).

Imaging was done with ultrasound machines mentioned in Fig. 1. Prior to imaging, a complete clinical breast examination was done by a referring physician or a resident radiologist. This examination was in accordance with institutional guidelines in each center. A variety of techniques, including superior-inferior, radial star-shaped sweeps, were used in patients based on the clinical judgment of the observing radiologist and institutional guidelines of each of the centers. All four quadrants were assessed, and angling of the probe or other necessary maneuvers were performed when necessary. All detected lesions were imaged in multiple axes; still, images were taken in the largest axis and considered for further analysis and quantification phases. All the images were saved, anonymized, and analyzed as bitmap files. The flow diagram of the study is presented in Fig. 2.

2.2. *Contouring of lesions*

Each lesion was segmented by two independent radiologists with at least 15 years of experience in breast ultrasound imaging. The Dice coefficient is calculated for each lesion to check the degree of agreement between the radiologists' contours (Fig. 2a). Contours with more similarity yield higher Dice coefficient values, and vice versa [23]. The binary contours and related images are fused for further feature extractions (Fig. 2b).

2.3. *Extraction of radiomics features*

Before feature extraction, 3σ normalization is done on each region to eliminate unwanted effects of using different equipment in the three centers. Overall, 242 radiomics features are extracted from each breast lesion using the MATLAB software (version R2019b, MathWorks Inc). These features belonged to one of the following groups: Histogram-based features (9 features: mean, variance, skewness, kurtosis, Perck% (k = 1, 10, 50, 90, and 99)); Autoregressive model (ARM, 5 features: Teta1, Teta2, Teta3, Teta4, and Sigma); Absolute gradient matrix (AGM, 5 features: mean, variance, skewness, kurtosis, and percentage of nonzero pixels (NonZeros)); Histogram of oriented gradients (HOG, 60 features: HOG features are calculated for 4, 8, 16, and 32 angular bins); Gabor (24 features: magnitudes of Gabor transform are determined for different Gaussian envelopes (4, 6, 8, 12, 16); Gray-level run-length matrix (GLRLM) (28 features: Gray-level and Run-length nonuniformity (GLevNonUni, RLNonUni), Long and short run emphasis (LngREmph, ShrtREmph), Normalized GLevNonUni and RLNonUni (MGLevNonUni, MRLNonUni), and Fraction; Gray-level co-occurrence matrix (GLCM) (44 features: angular second moment (ASM), sum of squares (SOSq), correlation (Correlat), contrast, inverse difference moment (IDiffM), Sum variance (SVrc), Difference entropy (DEntp), Difference variance (DVrc), Entropy, Sum entropy (SEntp), Sum average (SAvg). The GLCM, Gabor, and GLRLM features are calculated for directions (horizontal (H), 45°(Z), 90°(V), and 135°(N));

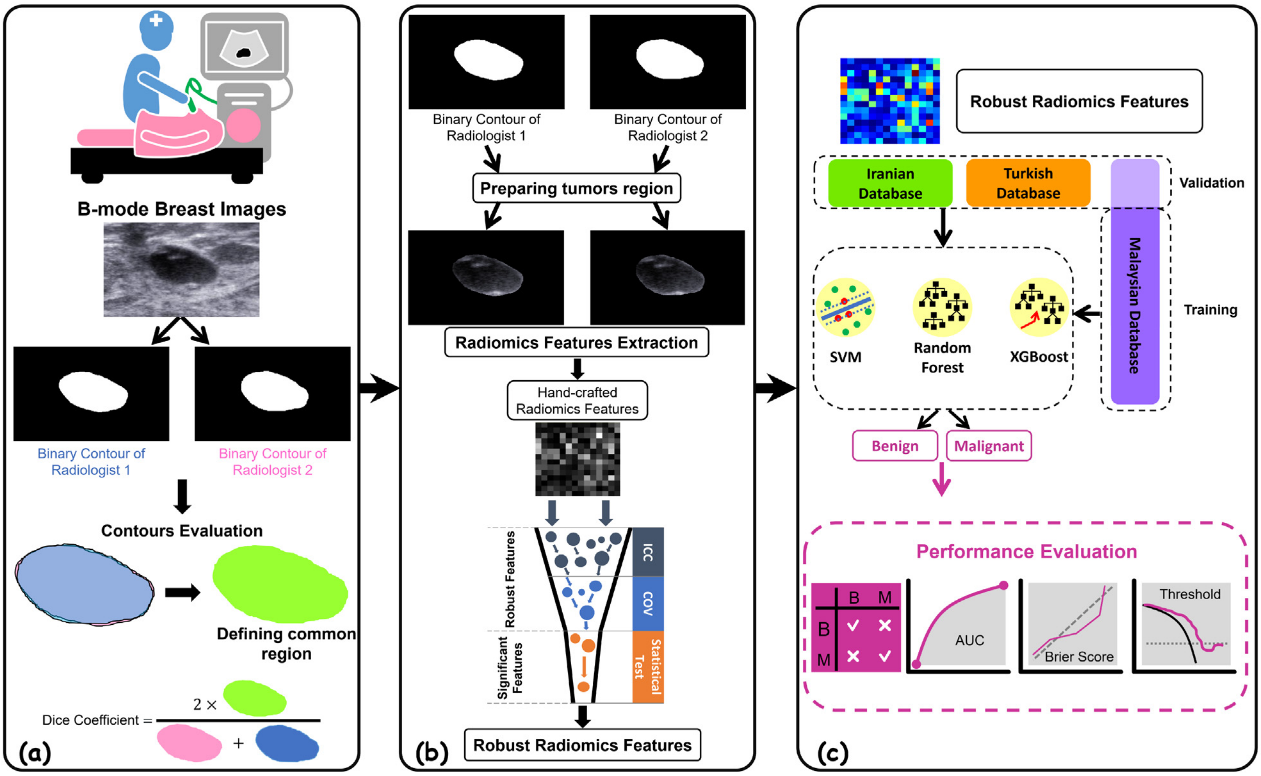


Fig. 2 – The flow diagram of the multistep process of image segmentation, feature extraction, model development, and performance evaluation utilized in the present study.

Wavelet (16 features: energy of wavelet coefficients for four-level decompositions; and local binary patterns (LPB, 51 features: LBP features are calculated for center-symmetric (Cs), transition (Tr), and over-complete (Oc) thresholding methods). Each of these features examines specific aspects of a given region of interest, including relationships between individual pixels, intensity changes (gradient) of pixels, the spatial location of pixels, and periodic textures [24,25].

2.4. *Statistical analysis*

Not all of the extracted radiomics features are included in the models due to the high dimensionality of data and the possibility of the existence of redundant features. A three-step process reduces dimensions and incorporates robust features into the machine-learning models (Fig. 2b). Intra-class correlation coefficients (ICCs) and coefficient of variation (COV) of radiomics features are determined. The features are considered robust if the ICC is greater than 0.95 and the COV is less than 0.4. COV is calculated by dividing the mean value (μ) by the standard deviation (σ) of each of the radiomics features ($\text{COV} = \mu/\sigma$). The independent sample *t*-test then evaluates features meeting both of these conditions to find significant robust radiomics features between the two groups. The significant features are considered for AI study. SPSS software (IBM SPSS Statistics, version 22.0) is used for statistical analysis, and a *p*-value < 0.05 is considered statistically significant.

2.5. *Machine-learning algorithms*

Robust significant radiomics features are fed into machine-learning algorithms, which aim to classify any given set of data into one of the two groups. The proposed method has two main phases of pre-processing and classification:

2.5.1. *Pre-processing*

The main task of the pre-processing phase is the transformation of feature values into a range of 0 to 1 using minimum–maximum normalization; this transformation prevents the machine-learning method from biasing toward features with a larger value [26].

2.5.2. *Classification*

A breast lesion is classified as benign or malignant in the classification phase. In this study, 3 classifiers, namely support vector machine classifier (SVM), random forest (RF), and gradient boosting (XGBoost), are employed to carry out the task of classification (Fig. 2c). The hyperparameters of each classifier were determined using the grid search method. In this study, SVM transfers the data into the Hilbert space via a second-order polynomial kernel with an automatic kernel coefficient to tackle the non-linearity of the data. Moreover, the L2 penalty with the regularization parameter of 1 regularizes the model for better performance on the test data [27,28]. RF classifier uses 15 decision trees as a base classifier with the Gini index as a splitting criterion. In this classifier, all of the features are considered to look for best split. Moreover, for better generalization, the maximum depth of base trees is limited to 5 [29]. XGBoost, as a Gradient-based ensemble method with 50 base learners, uses deviance loss with a learning rate of 0.1 for better performance. Moreover, the Friedman mean square error is employed to measure the quality of split [30].

2.5.3. *Performance evaluation*

To evaluate the performance of the proposed method, the classifiers are trained based on an 85 % of the Malaysian dataset (725 lesions: 422 benign, 303 malignant) and tested on the Malaysian, Iranian, and Turkish datasets. In other words, the performance of the proposed method has been evaluated both in 2 center-independent test datasets and in a center-dependent test dataset. Center-independent test datasets are Iranian and Turkish, while the center-dependent dataset is 15 % of the Malaysian dataset (128 lesions: 79 benign, 49 malignant) that was never included in the training procedure (Fig. 1). As the random sampling of 15 percent of the Malaysian dataset is stratified, the ratio between the numbers of samples in both benign and malignant classes is maintained, and the same datasets are considered for all classifiers. Moreover, 15 percent of the training set is used for the validation of machine learning models.

In order to quantify model performance, a variety of standard performance measures such as sensitivity (Sen), specificity (Spc), and accuracy (Acc) are utilized. Moreover, the ROC analysis is utilized to measure the area under the ROC curve (AUC) to estimate the overall performance of the proposed method. In addition, calibration curve analysis is performed to determine whether the predicted probability agrees with the real probability. This agreement is measured via the Brier score, which is defined in equation (1). In this equation, N, o_t, and f_t is the total number of predicting instances, the actual outcome of instance t, and the output probability of the model at instance t, respectively.

$$BrierScore = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \quad (1)$$

Furthermore, decision curve analysis (DCA) is performed to demonstrate the clinical practicability of the prediction models. Applying DCA determines whether using a predictor to make clinical decisions will provide benefit over alternative decision criteria, given a specified threshold probability (Fig. 2c). Python Sklearn packages are used for DCA and calibration curve analysis. All the scripts used in this study are made available via https://github.com/AliMedPhysics/Radiomics-based-Multi-Center-Breast-Cancer-Project. In addition, the quality of our radiomics study is evaluated using the radiomics quality score (RQS) at https://www.radiomics.world/rqs2.

3. Results

3.1. *Robust radiomics features*

Fig. 3a. indicates the agreement of the two radiologists in lesion contouring. The similarity of contours between the radiologists was regarded as excellent, as the Dice coefficient was 0.887 ± 0.056 (mean ± SD). Sample images of benign and malignant breast lesions are shown in Fig. 3b and 3f, respectively. The radiologists' contours for the benign and malignant lesions and their corresponded Dice coefficients are shown in Fig. 3c-e and Fig. 3g-i, respectively.

Of the 242 radiomics features extracted from our training dataset, 47 had a COV less than 0.4 in both benign and malignant lesions and also had an ICC > 0.95. The distribution of these features among benign and malignant groups is presented in Table 1. Of these features, only 33 had statistically significant differences between the 2 groups ($p < 0.05$). These features were labeled as robust features and were utilized for training our models. The feature reduction steps with the corresponding features are shown in Fig. 4.

3.2. *Results of AI models on multi-center datasets*

Internal validation was performed on 15 % of the data from the Malaysian center, and XGBoost achieved the highest performance with an AUC of 0.911. The lowest performance was recorded for the SVM-based model (AUC = 0.803). Similar results were seen in our first external validation dataset (Iran), where the best AUC was achieved by XGBoost (0.894). However, RF had a higher specificity than XGBoost, and the SVM model underperformed in this external dataset. For the third dataset (external validation 2, Turkey), the best performance was seen for RF with an AUC of 0.878, slightly higher than XGBoost (AUC = 0.875). However, the highest sensitivity

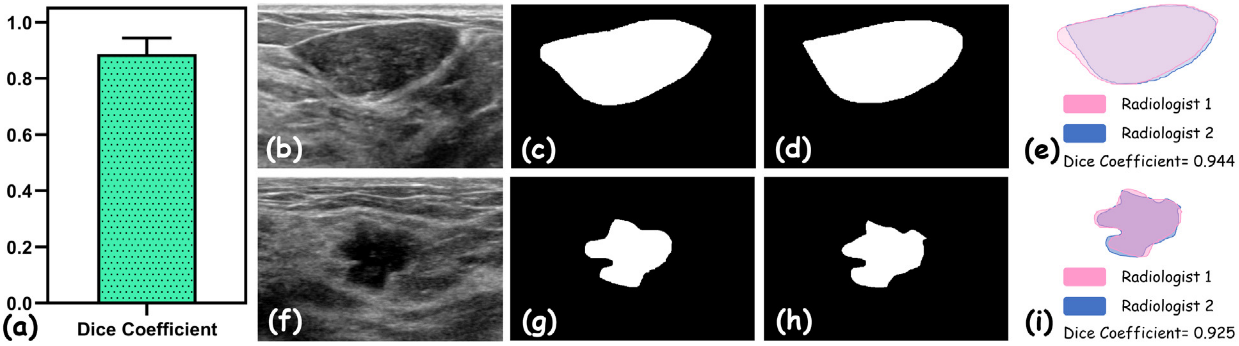
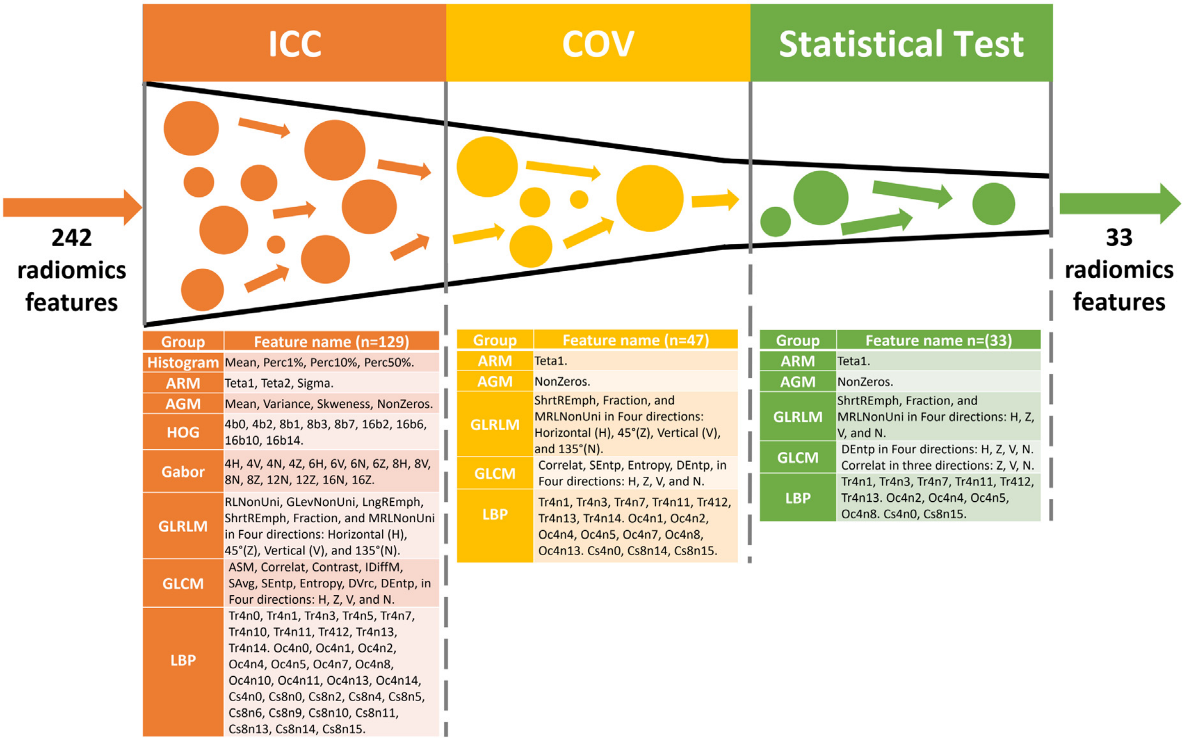


Fig. 3 – Results of radiologists' delineating based on Dice coefficient (a). Ultrasound sample images of benign (b) and malignant (f) breast lesions with their corresponding masks (c-d and g-h, respectively) are shown. Dice coefficient with calculation details for each benign (e) and malignant (i) lesions are provided.

Table 1 – List of robust radiomics features extracted from benign and malignant breast lesions.

Feature group	Feature name	Benign	Malignant	P-value	Feature group	Feature name	Benign	Malignant	P-value
Autoregressive Model	Teta1	0.751 ± 0.218	0.695 ± 0.198	<0.001	**Gray-level co-occurrence matrix**	NEntropy	3.315 ± 0.457	3.343 ± 0.421	0.377
Gradient	NonZeros	0.955 ± 0.115	0.938 ± 0.103	0.031		NDEntp	1.555 ± 0.186	1.468 ± 0.151	<0.001
Gray-level run-length matrix	HShrtREmph	0.918 ± 0.032	0.899 ± 0.030	<0.001		ZCorrelat	0.601 ± 0.181	0.683 ± 0.171	<0.001
	HFraction	0.845 ± 0.127	0.799 ± 0.118	<0.001		ZSEntp	2.166 ± 0.245	2.158 ± 0.220	0.677
	HMRLNonUni	0.812 ± 0.065	0.774 ± 0.058	<0.001		ZEntropy	3.315 ± 0.456	3.342 ± 0.420	0.394
	VShrtREmp	0.949 ± 0.028	0.938 ± 0.024	<0.001		ZDEntp	1.554 ± 0.184	1.467 ± 0.150	<0.001
	VFraction	0.894 ± 0.123	0.866 ± 0.111	0.001	**Local binary patterns**	Tr4n1	0.056 ± 0.010	0.058 ± 0.008	0.046
	VMRLNonUni	0.878 ± 0.059	0.854 ± 0.050	<0.001		Tr4n3	0.070 ± 0.025	0.063 ± 0.020	<0.001
	NShrtREmp	0.953 ± 0.027	0.943 ± 0.023	<0.001		Tr4n7	0.022 ± 0.008	0.020 ± 0.006	<0.001
	NFraction	0.903 ± 0.120	0.875 ± 0.109	0.001		Tr4n11	0.026 ± 0.008	0.024 ± 0.006	<0.001
	NMRLNonUni	0.887 ± 0.057	0.864 ± 0.048	<0.001		Tr4n12	0.068 ± 0.023	0.060 ± 0.018	<0.001
	ZShrtREmp	0.953 ± 0.027	0.943 ± 0.023	<0.001		Tr4n13	0.025 ± 0.009	0.021 ± 0.006	<0.001
	ZFraction	0.902 ± 0.120	0.875 ± 0.110	0.001		Tr4n14	0.022 ± 0.007	0.021 ± 0.005	0.238
	ZMRLNonUni	0.887 ± 0.057	0.864 ± 0.048	<0.001		Oc4n1	0.062 ± 0.015	0.061 ± 0.011	0.597
Gray-level co-occurrence matrix	HCorrelat	0.760 ± 0.193	0.778 ± 0.182	0.189		Oc4n2	0.066 ± 0.014	0.072 ± 0.010	<0.001
	HSEntp	2.190 ± 0.254	2.168 ± 0.230	0.197		Oc4n4	0.053 ± 0.013	0.059 ± 0.011	<0.001
	HEntropy	3.202 ± 0.434	3.174 ± 0.407	0.348		Oc4n5	0.016 ± 0.005	0.017 ± 0.004	0.022
	HDEntp	1.337 ± 0.181	1.253 ± 0.152	<0.001		Oc4n7	0.051 ± 0.019	0.0500 ± 0.01	0.857
	VCorrelat	0.621 ± 0.180	0.699 ± 0.174	<0.001		Oc4n8	0.065 ± 0.013	0.070 ± 0.010	<0.001
	VEntp	2.171 ± 0.246	2.162 ± 0.221	0.602		Oc4n13	0.051 ± 0.019	0.050 ± 0.015	0.521
	VEntropy	3.305 ± 0.452	3.326 ± 0.417	0.495		Cs4n0	0.288 ± 0.095	0.316 ± 0.083	<0.001
	VDEntp	1.531 ± 0.186	1.443 ± 0.149	<0.001		Cs8n14	0.067 ± 0.019	0.065 ± 0.015	0.130
	NCorrelat	0.594 ± 0.179	0.683 ± 0.171	<0.001		Cs8n15	0.059 ± 0.018	0.053 ± 0.015	<0.001
	NSEntp	2.165 ± 0.246	2.159 ± 0.220	0.717					



Group	Feature name (n=129)
Histogram	Mean, Perc1%, Perc10%, Perc50%.
ARM	Teta1, Teta2, Sigma.
AGM	Mean, Variance, Skweness, NonZeros.
HOG	4b0, 4b2, 8b1, 8b3, 8b7, 16b2, 16b6, 16b10, 16b14.
Gabor	4H, 4V, 4N, 4Z, 6H, 6V, 6N, 6Z, 8H, 8V, 8N, 8Z, 12N, 12Z, 16N, 16Z.
GLRLM	RLNonUni, GLevNonUni, LngREmph, ShrtREmph, Fraction, and MRLNonUni in Four directions: Horizontal (H), 45°(Z), Vertical (V), and 135°(N).
GLCM	ASM, Correlat, Contrast, IDiffM, SAvg, SEntp, Entropy, DVrc, DEntp, in Four directions: H, Z, V, and N.
LBP	Tr4n0, Tr4n1, Tr4n3, Tr4n5, Tr4n7, Tr4n10, Tr4n11, Tr412, Tr4n13, Tr4n14. Oc4n0, Oc4n1, Oc4n2, Oc4n4, Oc4n5, Oc4n7, Oc4n8, Oc4n10, Oc4n11, Oc4n13, Oc4n14, Cs4n0, Cs8n0, Cs8n2, Cs8n4, Cs8n5, Cs8n6, Cs8n9, Cs8n10, Cs8n11, Cs8n13, Cs8n14, Cs8n15.

Group	Feature name (n=47)
ARM	Teta1.
AGM	NonZeros.
GLRLM	ShrtREmph, Fraction, and MRLNonUni in Four directions: Horizontal (H), 45°(Z), Vertical (V), and 135°(N).
GLCM	Correlat, SEntp, Entropy, DEntp, in Four directions: H, Z, V, and N.
LBP	Tr4n1, Tr4n3, Tr4n7, Tr4n11, Tr412, Tr4n13, Tr4n14. Oc4n1, Oc4n2, Oc4n4, Oc4n5, Oc4n7, Oc4n8, Oc4n13. Cs4n0, Cs8n14, Cs8n15.

Group	Feature name n=(33)
ARM	Teta1.
AGM	NonZeros.
GLRLM	ShrtREmph, Fraction, and MRLNonUni in Four directions: H, Z, V, and N.
GLCM	DEntp in Four directions: H, Z, V, N. Correlat in three directions: Z, V, N.
LBP	Tr4n1, Tr4n3, Tr4n7, Tr4n11, Tr412, Tr4n13. Oc4n2, Oc4n4, Oc4n5, Oc4n8. Cs4n0, Cs8n15.

Fig. 4 – Feature reduction methods involved in this study. 242 Radiomics features were extracted from each lesion. 129 out of 242 features met the ICC > 0.95 criterion. 47 out of 129 features met the ICC + COV criteria (ICC > 0.95 & COV < 0.4). Finally, 33 features met all of the criteria (ICC > 0.95 & COV < 0.4 & P-value < 0.05 between the benign and malignant groups).

Table 2 – Performance of AI models in diagnosis of benign and malignant breast lesions for each validation datasets.

Dataset →	Internal Validation 1 (Malaysia)				External Validation 1 (Iran)				External Validation 2 (Turkey)				All Cohorts			
Model ↓	Sen	Spc	Acc	AUC	Sen	Spc	Acc	AUC	Sen	Spc	Acc	AUC	Sen	Spc	Acc	AUC
Gradient Boosting	0.918	0.899	0.906	0.911	0.862	0.917	0.888	0.894	0.960	0.788	0.826	0.875	0.903	0.867	0.884	0.890
Random Forest	0.816	0.899	0.867	0.860	0.732	0.963	0.840	0.848	0.880	0.879	0.879	0.878	0.793	0.916	0.859	0.860
SVM	0.735	0.873	0.820	0.803	0.683	0.661	0.672	0.673	0.813	0.525	0.649	0.670	0.733	0.672	0.700	0.695

Sen, sensitivity; Spc, specificity; Acc, accuracy; AUC, Area under the ROC curve.

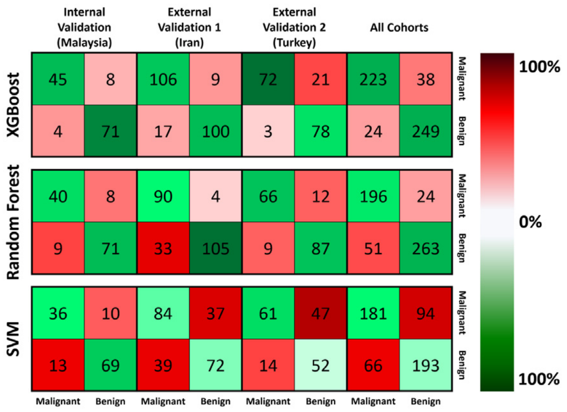
was achieved by the XGBoost-based model (96 %). The SVM-based model again underperformed (AUC = 0.670) compared to the two other models. When combining all three validation datasets, the highest performance was seen for the XGBoost-based model with an AUC of 0.890; more information is presented in Table 2 and Fig. 5. Corresponding ROC curves for each validation dataset and for all cohorts are presented in Fig. 6.

Calibration curves of the 3 AI models for the classification of benign and malignant breast lesions are shown in Fig. 7. The calibration curve of XGBoost classifier demonstrates higher agreement between observations and predictions in all of the testing cohorts (evident by the Brier score). A Low Brier score of the AI-based prediction models reveals high confidence in the outcome of the models. Moreover, the analysis of the decision curves is provided in Fig. 8. Fig. 8a-c demonstrates the DCA of internal validation, external validation 1, and external validation 2, respectively. According to Fig. 8d., a higher net benefit was achieved for the XGBoost classifier in diagnosing malignant breast lesions for all cohorts, with the threshold probability being within the range of 0.3 to 0.62. The RQS of our multi-center radiomics study was satisfactory at 59.09 %.

4. Discussion

The present study introduces significant robust features that can be used for multi-center diagnostic surveys and examines the diagnostic efficacy of machine-learning algorithms in classifying any solid non-cystic lesion as benign or malignant. The results indicated that XGBoost had the highest performance for all cohorts, with an AUC of 0.890.

There is an untapped potential in using AI-based– image trained models for breast lesion characterization, prognosis and therapy response determination, and even prediction of anatomic or histopathologic features of a lesion [31–33]. However, this potential has not been realized in clinical practice in many settings. Heterogeneity of the methods of the existing



		Internal Validation (Malaysia)		External Validation 1 (Iran)		External Validation 2 (Turkey)		All Cohorts	
		Malignant	Benign	Malignant	Benign	Malignant	Benign	Malignant	Benign
XGBoost	Malignant	45	8	106	9	72	21	223	38
	Benign	4	71	17	100	3	78	24	249
Random Forest	Malignant	40	8	90	4	66	12	196	24
	Benign	9	71	33	105	9	87	51	263
SVM	Malignant	36	10	84	37	61	47	181	94
	Benign	13	69	39	72	14	52	66	193

Fig. 5 – Confusion matrix of each AI model for all validation cohorts. The vertical and horizontal directions indicate true and predicted classes, respectively.

pieces of evidence and scarcity of models tested on external datasets is one of the main reasons limiting the generalizability of the findings of most current studies and thus limiting their clinical applicability. Table 3 presents a list of studies evaluating deep- or machine-learning algorithms with external validation datasets. Romeo et al. performed a study where an RF-based model was trained on ultrasound images from 135 breast lesions and then was tested on a different dataset from the same country. Their model incorporated ten radiomics features extracted from ultrasound breast images segmented by a single expert radiologist and achieved an AUC of 0.820 for an external dataset [34]. Our XGBoost-based model demonstrated a superior performance (0.890 vs 0.820). Unlike Romeo and their colleagues, we incorporated two contours for each lesion and additional statistical analysis to find robust radiomics features. We also included more radiomics features from different groups of features in our model, and the feature selection method differed considerably between the two studies. A study utilizing six different

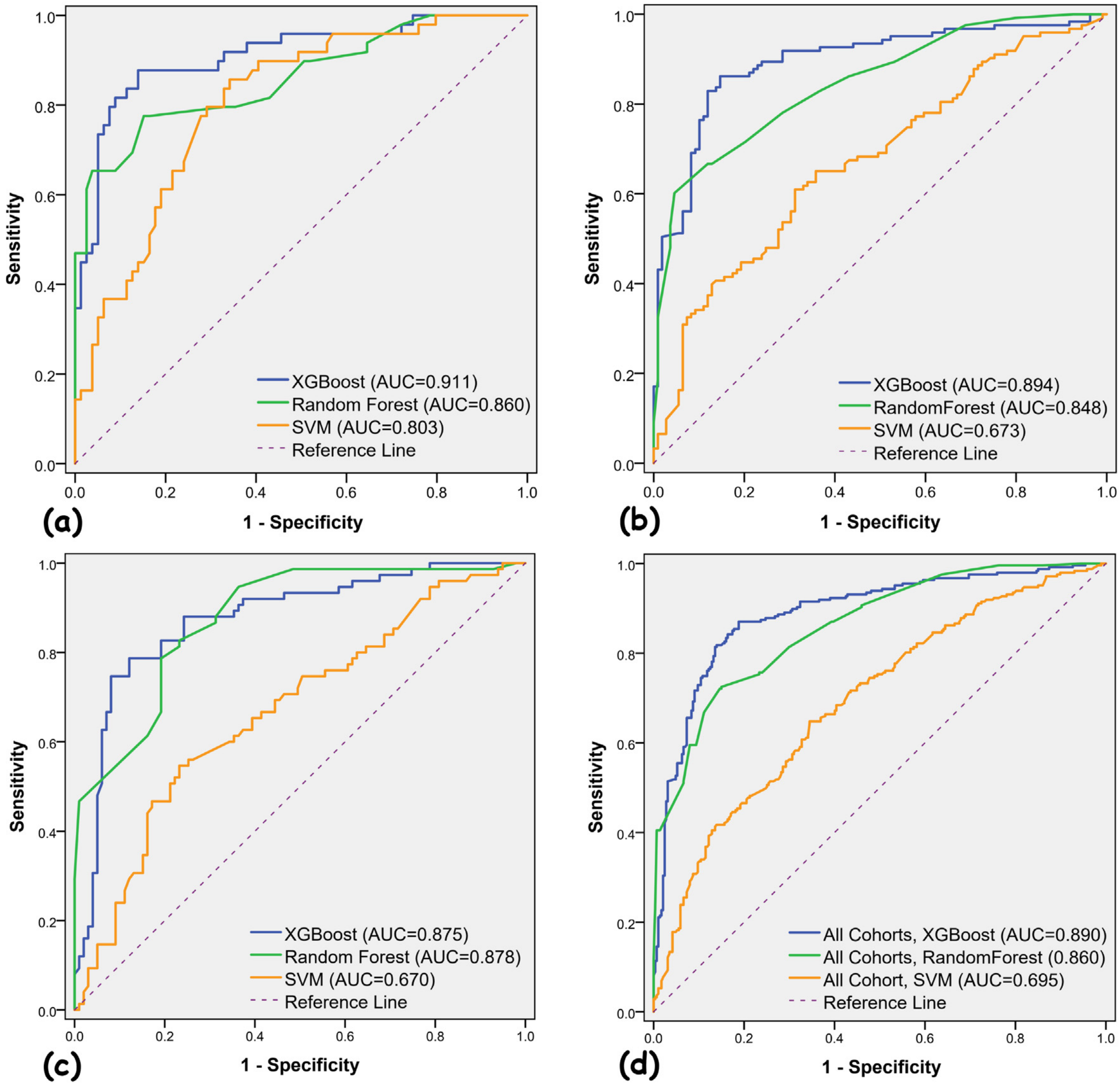


Fig. 6 – ROC curves of AI models in the classification of benign and malignant breast lesions for internal validation (Malaysia, a), external validation 1 (Iran, b), external validation 2 (Turkey, c), and all of the cohorts combined (d).

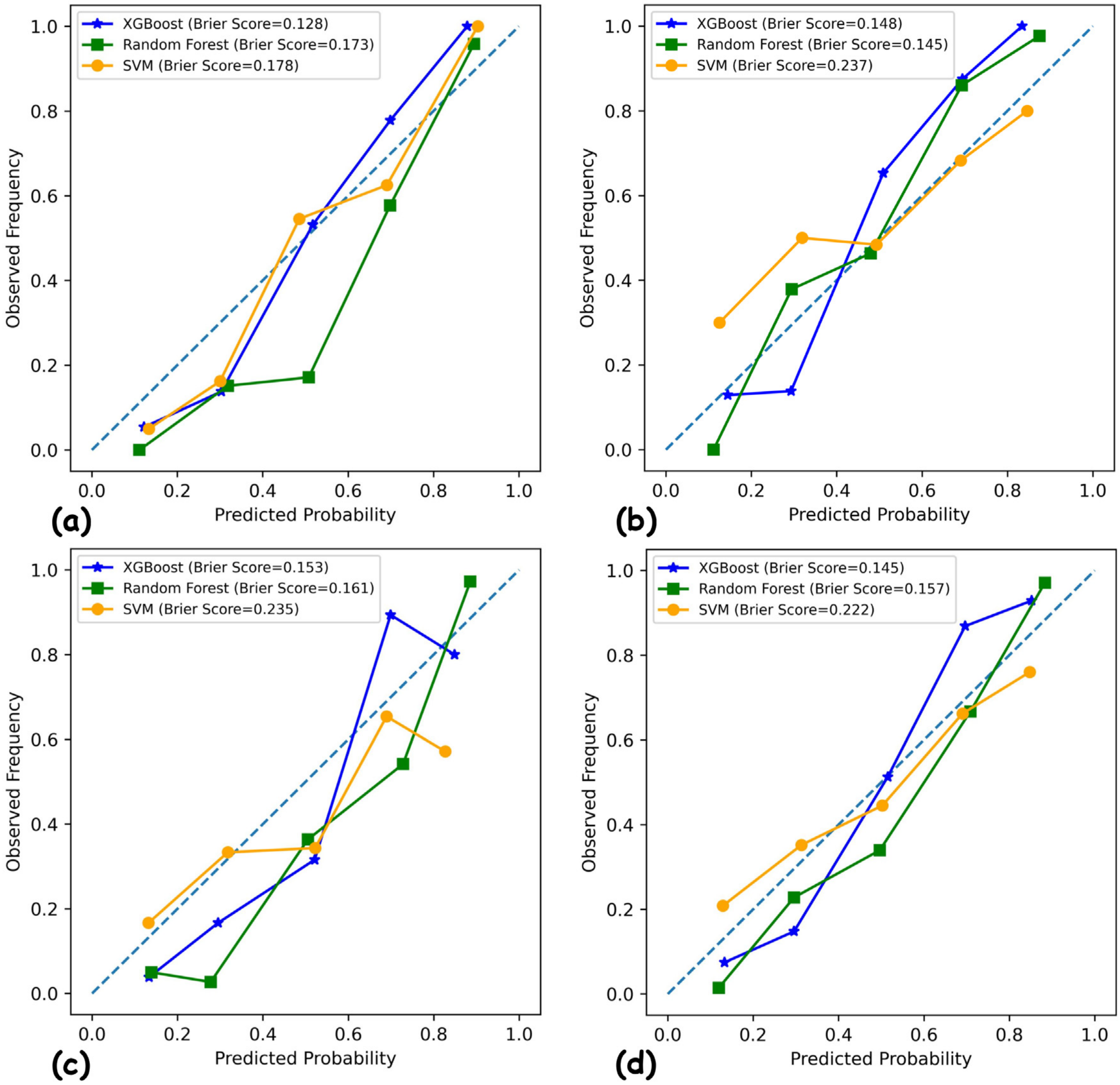


Fig. 7 – Calibration curves of AI models in the classification of benign and malignant breast lesions for internal validation (Malaysia, a), external validation 1 (Iran, b), external validation 2 (Turkey, c), and all of the cohorts combined (d).

machine-learning classifiers with an input of gross imaging characteristics was performed in China, where qualitative features of 1345 lesions were used to develop the models, and external validation was done using 1965 cases from 3 centers in the same country, China. The highest AUC was achieved for multilayer perceptron (0.775) in their test (external) dataset, which was inferior to our models in both external validation datasets (0.894, 0.878) [35]. The main reason our results are better than theirs is that we extracted radiomics features and determined robust features while they considered qualitative ultrasound features. Huo et al. entered breast cancer patients based on tumor staging, only including patients with T1 (or an equivalent stage). This clinical inclusion criterion may have wide-reaching effects on how models perform. Thus, results should be compared to studies with similar clinical inclusion criteria. Zhang et al. developed a deep-learning model to differentiate benign and malignant lesions and classify malignant lesions based on the existence of different molecular markers. Deep-learning models can extract infinite features from images and use the most efficient ones for classification. Although they may outperform conventional machine-learning models on extremely large data sets, their effectiveness in small to medium-sized datasets is a topic of discussion [36,37]. Their deep-learning model performed well in the external dataset and had an AUC of 0.900 (sensitivity of 81.7 % and specificity of 84.6 %). Another similar study by Kim et al. utilized a house-made deep-learning algorithm. Like the previous study, the algorithm achieved high sensitivity and specificity rates among three external validation datasets (sensitivity and specificity ranged between 76.16 % and 94.17 %, and between 81.67 % and 97.33 %, respectively) [38]. Interestingly, pretrained deep-learning algorithms also showed promising results in differentiating malignant and benign lesions. Kim et al. selected three well-known CNN algorithms and trained them on a medium-sized pool of lesions (1000). The results of external validation on three separate datasets showed that weakly-supervised deep-learning models provided excellent diagnostic performance (AUC: 0.86–0.90) that were non-inferior to the supervised model with either manual (AUC: 0.89–0.92) or

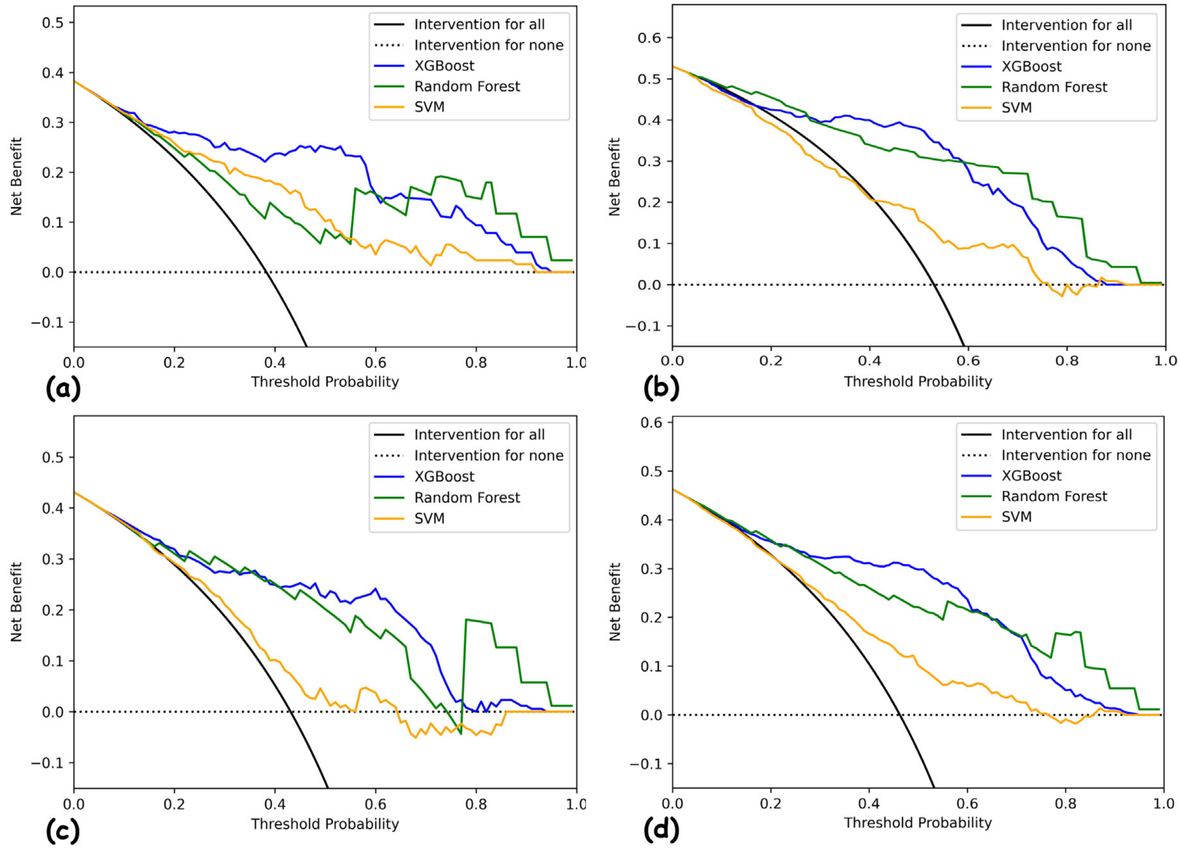


Fig. 8 – Decision curve analysis of AI models in the classification of benign and malignant breast lesions for internal validation (Malaysia, a), external validation 1 (Iran, b), external validation 2 (Turkey, c), and all of the cohorts combined (d).

automated annotation (AUC: 0.84–0.87). These results were particularly important because simpler deep-learning models could be used in clinical settings with minimal dependence on technical support [39]. The results of these 3 studies using deep-learning algorithms are comparable to ours, as we also witnessed a very suitable specificity using XGBoost. The proposed feature selection method plays a vital role in multi-center studies, demonstrated by the results of our machine-learning model, as it is comparable to the deep-learning models' results. Therefore, researchers who intend to establish a multi-center study protocol highly recommend considering the feature selection pipeline to determine statistically significant features and increase their models' generalizability.

Notably, all of the studies mentioned above performed external validation on one or more datasets which all belonged to a single country. In our study, validation was done on datasets from 3 different countries, Malaysia, Iran, and Turkey. This variation in data pools could hypothetically affect the algorithms' function, as it has been shown that breast density could vary considerably based on race and sub-race [40,41]. Furthermore, most of the focus on breast tissue heterogeneity among human populations has been directed to differences seen in African-Americans and sub-races of the Caucasian race. In contrast, the discrepancies between the subraces themselves and Asian people have rarely been studied [42,43]. Considering all of these, normalization was done, and just robust radiomics features were included in the model. Our model's performance in external datasets was comparable to that of the internal validation dataset, and a significant decline was not seen in their performances.

There has been much debate on the importance of robustness, reproducibility, and performance of classifier models since the introduction of radiomics features and their inevitable coupling to various classification systems, such as machine-learning algorithms. Currently, numerous guidelines point out criteria that should be considered before designing or performing a radiomics study [44]. Alongside these guidelines that should be strictly adhered to, there are other considerations when a model is being curated to perform on multiple external data sets [45,46]. In the present study, we aimed to determine a relevant clinical problem and utilize previously determined and agreed-upon solutions to make the study's methodology more rigorous and transparent [47]. However, there are limitations to our study, and our results should be viewed with attention to these limitations. We considered a small group of radiomics features (242 features) in this study. Furthermore, we had only 3 machine-learning classifiers. Further studies are needed to include other radiomics features and classifiers to make comprehensive results. Our classifier was trained on a single dataset and then externally validated using 2 datasets from 2 countries. Although we used a diverse set of data for patient selection in each of the data sets, differences in technical imaging specifications, the relevance of screening protocols, and differences witnessed among populations in regards to breast anatomy, density, and prevalence of breast cancer may nega-

Table 3 – List of studies aiming to classify breast lesions with external validation for deep or machine-learning algorithms trained on ultrasound images.

Study location	Training dataset	External validation dataset	Machine/deep learning method	Quantitative features extraction	External validation			Reference
					Sensitivity	Specificity	AUC	
Italy	135 lesions from a single center in Italy	66 lesions from a single center in Italy	Random Forest (RF) ensemble	10 features selected out of 520 features extracted	93 % (82–99)	57 % (34–89)	0.820 (0.700–0.900)	[34]
China	1345 lesions from a single center from China	1965 lesions from 3 centers in China	RF, LR, ET, SVM, Perceptron, XGboost	7 features	87 %	38 %	0.870	[35]
China	2,822 lesions from a single hospital in China	210 lesions from a single center in China	Xception convolutional neural network (DCNN)	Deep feature extraction	81.7 %	84.6 %	0.900	[36]
Republic of Korea	1000 lesions from a single center in the Republic of Korea	684 lesions from 3 separate centers in the Republic of Korea	Gaussian Pyramid with 4 c onvolution pathways	dynamic routing capsule network	94.17 %	81.67 %	0.900	[38]
Republic of Korea	1000 lesions from a single center in the Republic of Korea	200 lesions from a single center in the Republic of Korea	VGG16 (CNN) ResNet34 (CNN) GoogLeNet (CNN)	Deep feature extraction	91 % 78 % 88 %	72 % 80 % 76 %	0.910 0.890 0.920	[39]
The present study	725 lesions from a single center in Malaysia	534 lesions from 3 centers in Malaysia, Iran, and Turkey	Gradient boosting Random Forest SVM	33 robust features	90.3 % 79.3 % 73.3 %	86.7 % 91.6 % 67.2 %	0.890 0.860 0.695	NA

tively impact the extent to which our results could be generalized. Different institutional guidelines for care and management of suspected breast cancer patients may also affect the pool of included patients in different settings and result in asymmetries in the patients who may benefit from our algorithm, as our model was trained and tested on patients after they underwent specific selection criteria.

Large multi-center studies incorporating different lesions with various pathologies from patients of different races and ages would be of merit in deciphering the actual value of multiple machine-learning and deep-learning algorithms and their comparative efficiency. Furthermore, applying already programmed machine learning algorithms on different datasets from different countries with varying prevalence of breast cancer may facilitate the future development of newer, more robust models. The inclusion of clinical data, past medical records, and examination findings into machine learning models may also provide new opportunities for early and efficient detection of breast cancer.

5. Conclusion

The present study evaluated the performance of 3 highly regarded and hand-tuned machine-learning algorithms in differentiating benign and malignant breast lesions, using rigorous radiomics features as inputs. We found that the XGBoost algorithm had high diagnostic performance and a relatively unchanged diagnostic profile when used in external validation datasets. Our study provides evidence regarding the utility of externally evaluated machine-learning algorithms in differentiating breast lesions in clinical practice.

Ethical approval

Institutional Review Board (IRB) approval for data collection was obtained for each participating center:

1. Research Ethics Committees of Vice-Chancellor in Research Affairs - Shahid Beheshti University of Medical Sciences: IRB #IR.SBMU.RETECH.REC.1400.215.
2. The dataset from Malaysia was supported in part by the Ministry of Higher Education Malaysia FRGS /1/2019 SKK03/UM/01/1 (MOHE). The dataset by University of Malaya is protected and proprietary to internal institution. The dataset from Malaysia is approved by local institution ethics review board: IRB #MREC ID NO:2019822-7771.
3. University of Health Sciences, Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital, Istanbul, Turkey: IRB #252/2021.

CRediT authorship contribution statement

Hassan Homayoun: Formal analysis, Methodology, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Wai Yee Chan:** Data curation, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **Taha Yusuf Kuzan:** Data curation, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Wai Ling Leong:** Data curation, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **Kübra Murzoglu Altintoprak:** Data curation, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **Afshin Mohammadi:** Data curation, Methodology, Resources, Writing – original draft, Writing – review & editing. **Anushya Vijayananthan:** Data curation, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. **Kartini Rahmat:** Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Sook Sam Leong:** Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Mohammad Mirza-Aghazadeh-Attari:** Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Sajjad Ejtehadifar:** Data curation, Visualization, Writing – original draft, Writing – review & editing. **Fariborz Faeghi:** Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. **U. Rajendra Acharya:** Conceptualization, Investigation, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ali Abbasian Ardakani:** Conceptualization, Project administration, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Enzmann DR. Radiology's value chain. Radiology 2012;263 (1):243–52.

[2] Martin-Carreras T, Chen P-H. From data to value: how artificial intelligence augments the radiology business to create value. Semin Musculosk Radiol 2020;24:65–73.

[3] Dontchos BN, Yala A, Barzilay R, Xiang J, Lehman CD. External validation of a deep learning model for predicting mammographic breast density in routine clinical practice. Acad Radiol 2021;28:475–80.

[4] Chiwome L, Okojie OM, Rahman A, Javed F, Hamid P. Artificial intelligence: is it Armageddon for breast radiologists? Cureus 2020;12:e8923.

[5] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–10.

[6] Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. Clin Radiol 2019;74:357–66.

[7] Hickman SE, Baxter GC, Gilbert FJ. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. Br J Cancer 2021;125:15–22.

[8] Paravastu SS, Theng EH, Morris MA, Grayson P, Collins MT, Maass-Moreno R, et al. Artificial intelligence in vascular-PET: translational and clinical applications. PET Clin 2022;17:95–113.

[9] Piri R, Edenbrandt L, Larsson M, Enqvist O, Skovrup S, Iversen KK, et al. "Global" cardiac atherosclerotic burden assessed by artificial intelligence-based versus manual segmentation in

18F-sodium fluoride PET/CT scans: Head-to-head comparison. J Nucl Cardiol. 2021:1-9.

[10] Hansen JA, Naghavi-Behzad M, Gerke O, Baun C, Falch K, Duvnjak S, et al. Diagnosis of bone metastases in breast cancer: Lesion-based sensitivity of dual-time-point FDG-PET/CT compared to low-dose CT and bone scintigraphy. PloS One 2021;16:e0260066.

[11] Wallis M, Tarvidon A, Helbich T, Schreer I. Guidelines from the European Society of Breast Imaging for diagnostic interventional breast procedures. Eur Radiol 2007;17:581–8.

[12] Black E, Richmond R. Improving early detection of breast cancer in sub-Saharan Africa: why mammography may not be the way forward. Global Health 2019;15:3.

[13] Hill DA, Haas JS, Wellman R, Hubbard RA, Lee CI, Alford-Teaster J, et al. Utilization of breast cancer screening with magnetic resonance imaging in community practice. J Gen Intern Med 2018;33:275–83.

[14] Feig S. Cost-effectiveness of mammography, MRI, and ultrasonography for breast cancer screening. Radiol Clin North Am 2010;48:879–91.

[15] Rebolj M, Assi V, Brentnall A, Parmar D, Duffy SW. Addition of ultrasound to mammography in the case of dense breast tissue: systematic review and meta-analysis. Br J Cancer 2018;118:1559–70.

[16] Sood R, Rositch AF, Ambinder E, Pool K-L, Shakoor D, Pollack E, et al. Ultrasound for breast cancer detection in low-resource settings: systematic review and meta-analysis. Am Soc Clin Oncol 2018.

[17] Hooley RJ, Scoutt LM, Philpotts LE. Breast ultrasonography: state of the art. Radiology 2013;268:642–59.

[18] Pan H-B. The role of breast ultrasound in early cancer detection. J Med Ultrasound 2016;24:138–41.

[19] Shen Y, Shamout FE, Oliver JR, Witowski J, Kannan K, Park J, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat Commun 2021;12:1–13.

[20] Sechopoulos I, Mann RM. Stand-alone artificial intelligence - The future of breast cancer screening? Breast (Edinburgh, Scotland) 2020;49:254–60.

[21] Avanzo M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, et al. Machine and deep learning methods for radiomics. Med Phys 2020;47:e185–202.

[22] Harding-Theobald E, Louissaint J, Maraj B, Cuaresma E, Townsend W, Mendiratta-Lala M, et al. Systematic review: radiomics for the diagnosis and prognosis of hepatocellular carcinoma. Aliment Pharmacol Ther 2021;54:890–901.

[23] Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 2004;11:178–89.

[24] Abbasian Ardakani A, Bureau NJ, Ciaccio EJ, Acharya UR. Interpretation of radiomics features: a pictorial review. Comput Methods Programs Biomed 2022;215:106609.

[25] Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures. They are data. Radiology 2016;278:563–77.

[26] Singh D, Singh B. Investigating the impact of data normalization on classification performance. Appl Soft Comput 2020;97:105524.

[27] Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565–7.

[28] Meyer D, Leisch F, Hornik K. The support vector machine under test. Neurocomputing 2003;55:169–86.

[29] Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens 2005;26:217–22.

[30] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorob 2013;7:21.

[31] Arefan D, Hausler RM, Sumkin JH, Sun M, Wu S. Predicting cell invasion in breast tumor microenvironment from radiological imaging phenotypes. BMC Cancer 2021;21:370.

[32] Jiang M, Li C-L, Chen R-X, Tang S-C, Lv W-Z, Luo X-M, et al. Management of breast lesions seen on US images: dual-model radiomics including shear-wave elastography may match performance of expert radiologists. Eur J Radiol 2021;141:109781.

[33] Mao N, Yin P, Zhang H, Zhang K, Song X, Xing D, et al. Mammography-based radiomics for predicting the risk of breast cancer recurrence: a multicenter study. Br J Radiol 2021;94:20210348.

[34] Romeo V, Cuocolo R, Apolito R, Stanzione A, Ventimiglia A, Vitale A, et al. Clinical value of radiomics and machine learning in breast ultrasound: a multicenter study for differential diagnosis of benign and malignant lesions. Eur Radiol 2021;31:9511–9.

[35] Huo L, Tan Y, Wang S, Geng C, Li Y, Ma X, et al. Machine learning models to improve the differentiation between benign and malignant breast lesions on ultrasound: A multicenter external validation study. Cancer Manage Res 2021;13:3367–79.

[36] Zhang X, Li H, Wang C, Cheng W, Zhu Y, Li D, et al. Evaluating the accuracy of breast cancer and molecular subtype diagnosis by ultrasound image deep learning model. Front Oncol 2021;11:623506.

[37] Cheng PM, Montagnon E, Yamashita R, Pan I, Cadrin-Chênevert A, Perdigón Romero F, et al. Deep learning: an update for radiologists. Radiographics 2021;41:1427–45.

[38] Kim C, Kim WH, Kim HJ, Kim J. A multi-scale capsule network for improving diagnostic generalizability in breast cancer diagnosis using ultrasonography. In: International Workshop on PRedictive Intelligence in MEdicine. Springer; 2021. p. 181–91.

[39] Kim J, Kim HJ, Kim C, Lee JH, Kim KW, Park YM, et al. Deep learning-based breast cancer diagnosis at ultrasound: initial application of weakly-supervised algorithm without image annotation original research. 2021.

[40] Moore JX, Han Y, Appleton C, Colditz G, Toriola AT. Determinants of mammographic breast density by race among a large screening population. JNCI Cancer Spectrum 2020;4:pkaa010.

[41] Galukande M, Kiguli-Malwadde E. Mammographic breast density patterns among a group of women in sub Saharan Africa. Afr Health Sci 2012;12:422–5.

[42] Ellison-Loschmann L, McKenzie F, Highnam R, Cave A, Walker J, Jeffreys M, et al. Age and ethnic differences in volumetric breast density in New Zealand women: a cross-sectional study. PloS One 2013;8:e70217.

[43] El-Bastawissi AY, White E, Mandelson MT, Taplin S. Variation in mammographic breast density by race. Ann Epidemiol 2001;11:257–63.

[44] van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-“how-to” guide and critical reflection. Insights Imag 2020;11:91.

[45] Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J 2021;14:49–58.

[46] Snell KIE, Archer L, Ensor J, Bonnett LJ, Debray TPA, Phillips B, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. J Clin Epidemiol 2021;135:79–89.

[47] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749–62.