2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)
Mexico City, Mexico. September 11-13, 2019

A Performance Evaluation of Machine Learning Techniques for Breast Ultrasound Classification

Francisco A. González-Luna
CINVESTAV-IPN
Tamaulipas Campus
Ciudad Victoria, Mexico
fgonzalez@tamps.cinvestav.mx

Juanita Hernández-López
CINVESTAV-IPN
Tamaulipas Campus
Ciudad Victoria, Mexico
jhernandez@tamps.cinvestav.mx

Wilfrido Gómez-Flores
CINVESTAV-IPN
Tamaulipas Campus
Ciudad Victoria, Mexico
wgomez@tamps.cinvestav.mx

***Abstract*—In this paper, a performance comparison of seven machine learning (ML) approaches for classifying breast lesions on ultrasound is presented. From a dataset with 2032 cases (1341 benign and 691 malignant), 137 morphological and texture features were extracted, aiming to describe the BI-RADS lexicon for masses. Support vector machine (SVM), k-nearest neighbor (kNN), radial basis function network (RBFN), linear discriminant analysis (LDA), multinomial logistic regression (MLR), random forest (RF), and AdaBoost (Ada) were evaluated in terms of sensitivity (SEN), specificity (SPE), accuracy (ACC), and area under the ROC curve (AUC). The results revealed that LDA obtained the best classification performance with ACC=0.89, SEN=0.82, SPE=0.93, and AUC=0.95. Contrarily, the kNN obtained the lowest classification performance with ACC=0.87, SEN=0.76, SPE=0.93, and AUC=0.91. This results point out that the LDA classifier can be convenient to be used in a CAD system because it is simple to implement, and it does not require the tuning of hyperparameters.**

***Index Terms*—breast ultrasound; BI-RADS; tumor classification; supervised learning**

I. Introduction

Breast cancer is one of the most frequently diagnosed cancer and the leading cause of cancer death among women worldwide [1]. Early diagnosis is crucial for the survival of patients, where medical images are important sources of diagnostic information. In particular, breast ultrasound (BUS) is a coadjuvant technique to mammography (X-ray) in patients with palpable masses and normal or inconclusive mammogram findings. Also, BUS images can be useful to differentiate between benign and malignant tumors [2].

Computer-aided diagnosis (CAD) systems have emerged to perform image analysis to assist radiologists in image interpretation. Generally, the pipeline of a CAD system involves image preprocessing, lesion segmentation, feature extraction, and tumor classification [3]. In the literature, a plethora of approaches has been proposed to address each stage of a CAD system for BUS images [3, 4] .

When a CAD system analyzes a tumor, the feature extraction stage produces a feature vector whose entries are morphological and texture attributes derived from the Breast Imaging Reporting and Data System (BI-RADS) lexicon [5] to describe the shape, orientation, margin, echo pattern, and posterior feature of masses. Next, a machine learning (ML) approach classifies the tumor in one of the predefined classes, which are commonly benign and malignant classes [6].

It is necessary to select an ML approach that guarantees adequate generalization in new tumor cases. In this context, Shan et al. [5] assessed the classification performance of four ML techniques. Also, 10 features to quantify BI-RADS lexicon for masses were used. The random forest classifier attained the highest overall accuracy with 78.5%, which was obtained from a BUS dataset with 283 cases. Note that this overall accuracy can be improved by increasing the number of cases in the dataset, using a distinct set of features, and evaluating other ML approaches.

Therefore, aiming to improve the accuracy of tumor classification, in this paper, seven distinct ML techniques are evaluated, which are trained with 137 features extracted from a BUS dataset with 2032 cases.

II. Methods

A. Image dataset

The BUS dataset comprised 2032 BUS images provided by the National Cancer Institute (INCa) of Rio de Janeiro, Brazil. The INCa Research Ethics Committee had approved this study (protocol 38/2001). Patients were informed of the purpose of the study before consenting to participate. The images were collected from three ultrasound scanners: Logiq 6p (General Electric, Milwaukee, WI, USA), Logiq 5 (General Electric), and Sonoline Sienna (Siemens, Erlangen, Bavaria, Germany), with linear transducer arrays with frequencies between 7.5 and 12 MHz. The images were captured directly from the 8-bit video signal (i.e., 256 gray levels) and saved in TIFF

Fondo SEP-Cinvestav 2018 grant No. 145 and CONACyT scholar grant No. 490240

978-1-7281-4840-3/19/$31.00 ©2019 IEEE

format. All the cases were biopsy-proven: 1341 corresponded to benign tumors and 691 to malignant tumors.

A senior radiologist manually segmented each breast tumor in the dataset. These manual segmentations were used to locate the region of interest for further extraction of morphological and texture features. It is expected that the outlining of an expert contains relevant pixels that enclose the actual tumor shape. However, in a practical CAD system, an accurate computerized segmentation method can be used. Figure 1 shows some examples of BUS images used in this study.

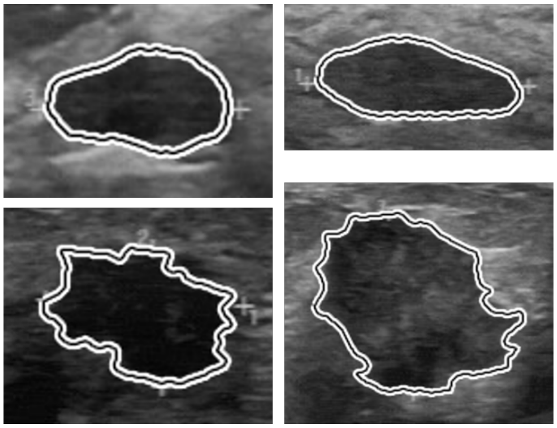

Fig. 1. Example of BUS images with manual segmentation. Top row are benign tumors and bottom row are malignant tumors.

B. Proposed approach

Figure 2 shows a block diagram of the proposed approach. From the region of interest containing a lesion, 137 morphological and texture features are extracted. These features are divided into five categories according to the BI-RADS lexicon for masses: shape, margin, orientation, echo pattern, and posterior feature. Next, these features are used to train a classifier. The output is a class label that indicates the probable nature of the lesion, that is, benign or malignant, which means, in terms of clinical recommendation, "no-biopsy" or "biopsy", respectively [7].

1) BI-RADS features: To consider the entire BI-RADS lexicon for masses, 41 morphological features, computed from the binary lesion shape, and 96 texture features, extracted from the intensity data in the region of interest, were calculated. Morphological features describe the shape, orientation, and margin of the lesion, whereas texture features characterize the lesion's boundary, echo pattern, and posterior feature [7]. In Table I are summarized the features extracted in this study. Moreover, the numerical implementation of these features can be found in [8].

2) Machine learning approaches: Seven ML approaches were trained for classifying breast lesions into benign and malignant classes. The ML approaches included the following methods [24, 25]:

- **Support vector machine** (SVM) is a large margin classifier which attempts to separate instances of different classes with the maximum margin hyperplane. The Gaussian kernel is used to create non-linear decision boundaries between classes.
- **k-nearest neighbors** (kNN) algorithm identifies the k instances from the training set that are closest to the test instance, which is classified according to the majority class among the k instances.
- **Radial basis function network** (RBFN) is an artificial neural network with three layers: input, hidden, and output. It uses radial basis functions (RBF) as activation functions in the hidden layer. The output of the network is a linear combination of RBF's responses of the inputs and output neuron weights.
- **Linear discriminant analysis** (LDA) maximizes the ratio "between-class scatter" to "within-class scatter", that is, making the distance between centers of different classes large while keeping the variance within each class small.
- **Multinomial logistic regression** (MLR) uses a logistic function to model a binary dependent variable; hence, it returns a probability score between 0 and 1. The parameters of MLR are learned through gradient descent to minimize an error cost function.
- **Random forest** (RF) is an ensemble method in which hundred of decision trees are trained independently. RF is based on bootstrap aggregating to obtain the data subsets to train each base classifier. Also, during the construction of a decision tree, RF selects a subset of features at each step of the split. The classification is performed by majority vote.
- **AdaBoost** (Ada) is a boosting algorithm which converts a set of weak classifiers into a strong one. The output of weak learners is combined into a weighted sum that represents the final output of the boosted classifier.

C. Experimental setup

The experimentation is focused on evaluating the classification performance of the seven above-mentioned ML approaches. First, before the training procedure, all the features were rescaled to the range $[-1, 1]$ by softmax normalization to reduce the influence of extreme feature values [26].

To evaluate the classification performance of each ML technique, k-fold cross-validation method partitioned the original data into k equal-size subsets, and then k runs of training-tests were performed. The k-fold cross-validation can be repeated t times to reduce the influence of randomness introduced by data split. Herein, $k = 10$ and $t = 10$ [24].

It is worth mentioning that SVM, RBFN and kNN classifiers required adjusting hyperparameters. Hence, the optimal hyperparameters were found by means of k-fold cross-validation (with $k = 5$) and a grid search [27]. For the SVM, the penalization parameter C and the Gaussian kernel bandwidth γ were searched in the range $C = [2^{-5}, 2^{-4}, \ldots, 2^{15}]$ and $\gamma = [2^{-15}, 2^{-14}, \ldots, 2^{3}]$, respectively. For the RBFN, the number of hidden units was searched in the range $[5, 100]$ with steps of 5. Finally, for kNN, the number of k nearest neighbors involved $[1, 3, 5, 7]$.

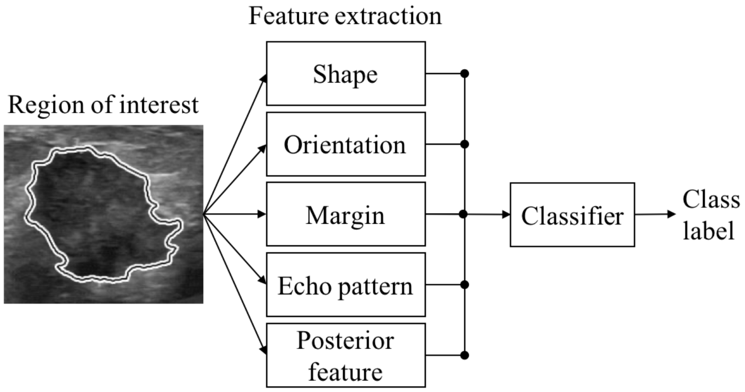

Fig. 2. The organization of the proposed approach.

TABLE I
MORPHOLOGICAL (M) AND TEXTURE (T) FEATURES USED TO DESCRIBE THE BI-RADS LEXICON FOR MASSES.

BI-RADS lexicon	Number of features	Quantitative descriptors	References
Shape (M)	26	Solidity, Normalized residual value, Convexity, Roundness, Circularity, Compactness, Eccentricity, Elongatedness, Form factor, Long-short axis ratio, Elliptic-normalized circumference, Elliptic-normalized skeleton, Area difference with equivalent ellipse, Extent, Shape class, Symmetry area, Average proportional distance, Maximun proportional distance, Pearson correlation coefficient, Mean squared error	[8–15]
Orientation (M)	2	Length-to-width ratio, Orientation	[15, 16]
Margin (M/T)	15	Spiculation, Lobule Index, Number of depressions, Number of undulations, Number of protuberances and depressions, External area ratio, Crossing, Average distance, Maximun distance, Roughness, Power espectral entropy, Fractal dimension, Abrupt Interface, Normalized radial gradient	[13, 15–19]
Echo pattern (T)	90	Gray-Level Concurrence Matrix (GLCM) features (Energy, Entropy, Correlation, Inverse difference moment normalized, Inertia, Cluster shade, Cluster prominence, Autocorrelation, Dissimilarity, Homogeneity, Maximun probability), Auto-Mutual information	[20–22]
Posterior feature (T)	4	Posterior acoustic feature, Minimun side diference, Standard deviation of posterior acoustic, Mean of the posterior acoustic and adjacent area	[15, 16, 23]

The classification performance was evaluated by the area under the ROC curve at cut-off point (AUC_c) [28]:

$$\mathrm{AUC}_c = \frac{1}{2}\left(\mathrm{SEN} + \mathrm{SPE}\right), \tag{1}$$

where the sensitivity (SEN) and specificity (SPE) are defined as

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad \text{and} \quad \mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{2}$$

where TP, FP, TN, and FN are calculated from observations classified by some ML approach, and represent the entries of the confusion matrix shown in Table II. The classification accuracy is computed as

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{N}, \tag{3}$$

where N is the sum of all the elements in the confusion matrix shown in Table II. The indices AUC_c, ACC, SEN, and SPE should tend toward unity to indicate an adequate classification performance.

TABLE II
CONFUSION MATRIX FOR BINARY CLASSIFICATION.

Class	Classified as malignant	Classified as benign
Malignant	True positive (TP)	False negative (FN)
Benign	False positive (FP)	True negative (TN)

The Kruskal-Wallis test ($\alpha = 0.05$) with the Bonferroni correction were used to compare the classification performance of distinct ML approaches statistically [29, 30].

III. RESULTS

The classification performance results are summarized in Table III, in which the seven ML approaches are compared by their mean and standard deviation values, calculated from 10-times 10-folds cross-validation.

In general, LDA obtained the best classification performance with $\mathrm{ACC} = 0.89$, $\mathrm{SEN} = 0.82$, $\mathrm{SPE} = 0.93$, and $\mathrm{AUC}_c = 0.88$. The lowest sensitivity was obtained by the kNN classifier with $\mathrm{SEN} = 0.76$, which negatively impacted on the overall classification rate; hence, kNN reached

TABLE III
CLASSIFICATION PERFORMANCE OF ML APPROACHES.

Classifier	ACC	SEN	SPE	AUC$_c$
Ada	0.89(0.02)	0.79(0.04)	0.94(0.02)	0.86(0.02)
kNN	0.87(0.02)	0.76(0.04)	0.93(0.02)	0.84(0.02)
LDA	0.89(0.02)	0.82(0.04)	0.93(0.02)	0.88(0.02)
MLR	0.89(0.02)	0.81(0.04)	0.92(0.02)	0.87(0.02)
RBFN	0.88(0.02)	0.77(0.04)	0.93(0.02)	0.85(0.02)
RF	0.88(0.02)	0.80(0.04)	0.93(0.02)	0.86(0.02)
SVM	0.89(0.02)	0.80(0.05)	0.93(0.02)	0.87(0.03)

TABLE IV
PAIRWISE STATISTICAL COMPARISON WITH KRUSKAL-WALLIS WITH BONFERRONI CORRECTION OF ML APPROACHES.

Classifier A	Classifier B	ACC	SEN	SPE	AUC$_c$
Ada	kNN	+	+	+	+
Ada	LDA	=	-	=	=
Ada	MLR	=	=	+	=
Ada	RBFN	+	=	=	+
Ada	RF	=	=	=	=
Ada	SVM	=	=	=	=
kNN	LDA	-	-	=	-
kNN	MLR	-	-	=	-
kNN	RBFN	-	=	=	=
kNN	RF	-	-	=	-
kNN	SVM	-	-	=	-
LDA	MLR	=	=	+	=
LDA	RBFN	+	+	=	+
LDA	RF	+	=	=	+
LDA	SVM	=	=	=	=
MLR	RBFN	=	+	=	+
MLR	RF	=	=	=	=
MLR	SVM	=	=	=	=
RBFN	RF	=	-	=	-
RBFN	SVM	=	-	=	-
RF	SVM	=	=	=	=

the lowest accuracy and AUC$_c$ values with 0.87 and 0.84, respectively.

In addition, all the ML approaches obtained similar values of specificity. For all classifiers the range was $0.92 \leq \text{SPE} \leq 0.94$. On the other hand, the values of sensitivity among ML techniques presented more variation, that is, $0.76 \leq \text{SEN} \leq 0.82$.

In agreement with the pairwise statistical comparison in Table IV, LDA and SVM performed statistically similar regarding all the performance indices, although the former requires a less computational cost to be trained because the SVM requires finding the optimal C and γ hyperparameters. Besides, LDA tends to statistically outperform the rest of ML techniques ($p < 0.001$).

Figure 3 shows the ROC curves of each ML approach. It is observed that LDA obtained the highest AUC value with 0.95, which confirms the findings in Table III. Nevertheless, classifiers such as SVM, RF, and MLR presented acceptable AUC values of 0.94.

IV. DISCUSSION AND CONCLUSIONS

Early diagnosis of breast cancer is crucial for the survival of patients, making an important task seeking a CAD that implements an ML method with the ability to separate between benign and cancer cases efficiently. Hence, in this work, a classification performance comparison between seven ML approaches was presented.

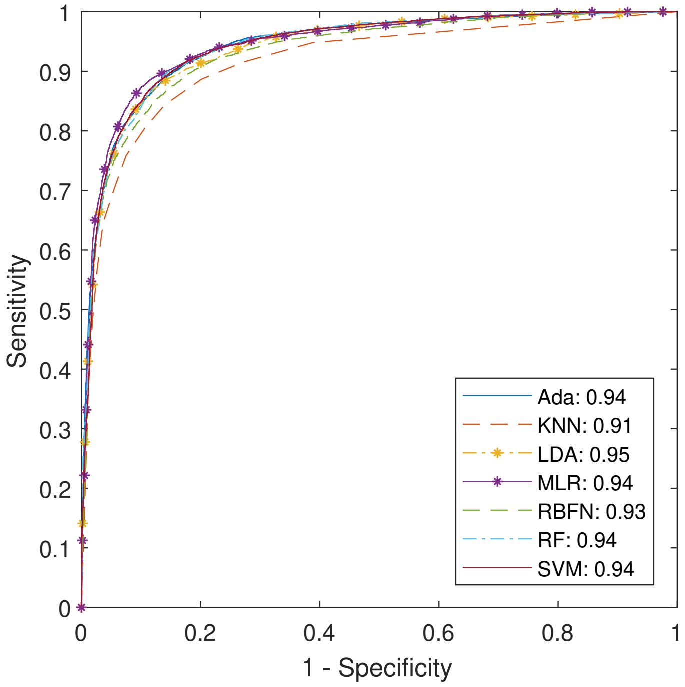

Fig. 3. ROC curves of ML approaches.

In the work of Shan et al. [5], four ML approaches, and 10 features were evaluated, by using a dataset with 283 BUS cases. The authors concluded that distinct classifiers have different performance in the classification of BUS images in benign and malignant classes. Herein, we extended the evaluation to seven ML approaches and 137 morphological and texture features. Also, a dataset with 2032 BUS cases was considered.

In general, LDA obtained the best classification performance and it did not present statistically significant differences with SVM. However, in practice, the LDA classifier can be convenient to be used in a CAD system because it is simple to implement and it does not require the tuning of hyperparameters (such as SVM and RBFN), which reduces the computational cost of training models.

It is notable that the specificity tends to reach values higher than 0.92 for all the evaluated ML methods. Nevertheless, the highest sensitivity was 0.82 attained by the LDA classifier. The trade-off between sensitivity and specificity can be improved by increasing the number of malignant cases aiming to obtain a more balanced dataset. Also, it can be evaluated distinct feature selection techniques to get a subset of relevant features to increase the classification performance.

It is worth mentioning that in the work of Shan et al. [5], the SVM reached the highest AUC value with 0.84, while the RF attained the highest accuracy value with 0.78. However, in this study, the SVM reached an AUC of 0.94, and the RF attained an accuracy of 0.88. Therefore, increasing the number of features as well as the number of cases positively impacted on the classification performance.

The experimental results confirmed the conclusion of Shan et al. [5], that is, using distinct ML approaches regarding the same feature set, different classification performances are obtained. Therefore, when a CAD system is being developed, it is recommended to determine the best ML approach for a specific feature space.

Future work includes the exploration of feature selection techniques for searching the optimal feature subset for each ML method. Also, because the current tendency of CAD systems is to use BI-RADS features on the description of the lesions, it could be convenient to provide information about the criteria that an ML technique used for making classification, that is, modeling the probability of malignancy of each BI-RADS lexicon.

Acknowledgment

This research was supported by a Fondo SEP-Cinvestav 2018 grant (No. 145) and a National Council of Science and Technology (CONACyT, Mexico) research scholar grant (No. 490240).

References

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *International Journal of Cancer*, vol. 136, no. 5, pp. E359–E386, 2015.

[2] A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: use of sonography to distinguish between benign and malignant lesions," *Radiology*, vol. 196, no. 1, pp. 123–134, 1995.

[3] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognition*, vol. 43, pp. 299–317, 2010.

[4] Q. Huang, Y. Luo, and Q. Zhang, "Breast ultrasound image segmentation: a survey," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–15, 2017.

[5] J. Shan, S. K. Alam, B. Garra, Y. Zhang, and T. Ahmed, "Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods," *Ultrasound in Medicine & Biology*, vol. 42, no. 4, pp. 980 – 988, 2016.

[6] N. I. Yassin, S. Omran, E. M. E. Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 156, pp. 25 – 45, 2018.

[7] A. Rodríguez, W. Gómez, and W. C. Albuquerque Pereira, "A computer-aided diagnosis system for breast ultrasound based on weighted BI-RADS classes," *Computer Methods and Programs in Biomedicine*, vol. 153, pp. 33 – 40, 2018.

[8] A. Rodriguez, W. Gomez, and W. C. Albuquerque Pereira, "BUSAT: A MATLAB toolbox for breast ultrasound image analysis," in *Pattern Recognition: 9th Mexican Conference, MCPR 2017, Proceedings*. LNCS 10267, 2017, pp. 268–277.

[9] Y.-L. Huang, D.-R. Chen, Y.-R. Jiang, S.-J. Kuo, H.-K. Wu, and W. K. Moon, "Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound," *Ultrasound in Obstetrics and Gynecology*, vol. 32, no. 4, pp. 565–572, 2008.

[10] A. V. Alvarenga, A. F. C. Infantosi, W. C. A. Pereira, and C. M. Azevedo, "Assessing the performance of morphological parameters in distinguishing breast tumors on ultrasound images," *Medical Engineering & Physics*, vol. 32, no. 1, pp. 49–56, 2010.

[11] B. Surendiran and A. Vadivel, "Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer," *International Journal of Medical Engineering and Informatics*, vol. 4, no. 1, pp. 36–54, 2012.

[12] F. Pak, H. R. Kanan, and A. Alikhassi, "Breast cancer detection and classification in digital mammography based on non-subsampled contourlet transform (NSCT) and super resolution," *Computer Methods and Programs in Biomedicine*, vol. 122, no. 2, pp. 89–107, 2015.

[13] C.-M. Chen, Y.-H. Chou, K.-C. Han, G.-S. Hung, C.-M. Tiu, H.-J. Chiou, and S.-Y. Chiou, "Breast lesions on sonograms: Computer-aided diagnosis with nearly setting-independent features and artificial neural networks," *Radiology*, vol. 226, no. 2, pp. 504–514, 2003.

[14] J. Shan, S. K. Alam, B. Garra, Y. Zhang, and T. Ahmed, "Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods," *Ultrasound in Medicine & Biology*, vol. 42, no. 4, pp. 980–988, 2016.

[15] W.-C. Shen, R.-F. Chang, W. K. Moon, Y.-H. Chou, and C.-S. Huang, "Breast ultrasound computer-aided diagnosis using BI-RADS features," *Academic Radiology*, vol. 14, no. 8, pp. 928–939, 2007.

[16] K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Computerized diagnosis of breast lesions on ultrasound," *Medical Physics*, vol. 29, no. 2, pp. 157–164, 2002.

[17] Y. Su, "Automatic detection and classification of breast tumors in ultrasonic images using texture and morphological features," *The Open Medical Informatics Journal*, vol. 5, no. 1, pp. 26–37, 2011.

[18] Y.-H. Chou, C.-M. Tiu, G.-S. Hung, S.-C. Wu, T. Y. Chang, and H. K. Chiang, "Stepwise logistic regression analysis of tumor contour features for breast ultrasound diagnosis," *Ultrasound in Medicine & Biology*, vol. 27, no. 11, pp. 1493–1498, 2001.

[19] R. M. Rangayyan and T. M. Nguyen, "Fractal analysis of contours of breast masses in mammograms," *Journal of Digital Imaging*, vol. 20, no. 3, pp. 223–237, 2006.

[20] M. Masotti, "A ranklet-based image representation for mass classification in digital mammograms," *Medical Physics*, vol. 33, no. 10, pp. 3951–3961, 2006.

[21] M.-C. Yang, W. K. Moon, Y.-C. F. Wang, M. S. Bae, C.-S. Huang, J.-H. Chen, and R.-F. Chang, "Robust texture analysis using multi-resolution gray-scale invariant features for breast sonographic tumor diagnosis," *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2262–2273, 2013.

[22] W. Gómez-Flores, A. Rodríguez-Cristerna, and W. C. de Albuquerque Pereira, "Texture analysis based on auto-mutual information for classifying breast lesions with ultrasound," *Ultrasound in Medicine & Biology*, vol. In Press, 2019.

[23] M. Qiao, Y. Hu, Y. Guo, Y. Wang, and J. Yu, "Breast tumor classification based on a computerized breast imaging reporting and data system feature system," *Journal of Ultrasound in Medicine*, vol. 37, no. 2, pp. 403–415, 2017.

[24] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC Press, 2012.

[25] R. Duda, P. Hart, and D. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[26] K. L. Priddy and P. E. Keller, *Artificial Neural Networks: An Introduction (SPIE Tutorial Texts in Optical Engineering, Vol. TT68)*. SPIE-International Society for Optical Engineering, 2005.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[28] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427 – 437, 2009.

[29] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*. Florida, USA: Chapman & Hall/CRC Press, 2011.

[30] H. Abdi, "The Bonferonni and Sidak corrections for multiple comparisons," in *Encyclopedia of Measurement and Statistics*, N. Salkind, Ed. Thousand Oaks, USA: SAGE Publications, Inc, 2007, pp. 103–107.