Breast Cancer Diagnosis Using Image Processing and Machine Learning for Elastography Images

Mohamed Adel[1,2], Ahmed Kotb[3], Omar Farag[3], M. Saeed Darweesh[1,4], Hassan Mostafa[4,5]
[1]*Institute of Aviation Engineering and Technology, Giza, Egypt*
[2]*National Institute of Laser Enhanced Sciences (NILES), Cairo University, Cairo, Egypt*
[3]*Germany University in Cairo (GUC), Cairo, Egypt*
[4]*University of Science and Technology, Nanotechnology Program, Zewail City of Science and Technology, October Gardens, 6th of October, Giza 12578, Egypt*
[5]*Faculty of Engineering, Cairo University, Giza, Egypt*

***Abstract*— As a trending medical imaging technique, Elastography and B-mode (ultrasound) are combined as a diagnostic tool to differentiate between benign and malignant breast lesions based on their stiffness and geometric properties. Image processing techniques are applied to the resulting images for feature extraction. Data preprocessing methods and principal component analysis (PCA) as a dimensionality reduction technique are applied to the dataset. In this paper, supervised learning algorithm "support vector machine (SVM)" is used for the classification of combined elastogram and B-mode images. Model validation is performed with K-fold cross-validation to ensure the generalization of the algorithm. Accuracy, confusion matrix, and logistic loss are then evaluated for the used algorithm. The maximum classification accuracy is 94.12% when using SVM with radial basis function (RBF) kernel.**

Keywords—Breast Cancer, Elastography, Image Processing, Principle component analysis, Support Vector Machine (SVM).

I. INTRODUCTION

Breast cancer constitutes a significant threat on women health and is considered the second leading cause of their death [1]. Breast cancer is a result of an abnormal behavior in the functionality of the normal breast cells. Therefore, breast cells tend to grow uncontrollably forming a tumor which can be felt as a lump in the breast [2].

Early diagnosis of breast cancer is proved to reduce the risks of death by providing a better chance of identifying a suitable treatment. In general, palpation, ultrasound and mammography are the most common ways of diagnosis. However, ultrasound elastography is currently playing a vital role in the process of breast cancer diagnosis. Computer-aided diagnosis using a combination of ultrasound (B-mode) and elastography images shows a noticeable superiority over other digital imaging techniques because of its accurate classification of lesions [3-4].

Machine learning makes use of mathematical and statistical models to learn from data. Machine learning finds an important role in biomedical applications in which accuracy of measurements is a crucial factor. Subsequently machine learning algorithms can help diagnose breast cancer at its early stages. Machine learning tools can determine most predicative features from complex and noisy datasets [5]. As a result, false negative and false positive decisions could be significantly reduced which yields better classification accuracy [6].

The paper is organized as follows: Section II gives descriptions of the materials and methods which are used. Section III investigates the performance metrics in detail. Simulation results and comparison are given in section IV. Finally, the whole work is concluded in section V.

II. MATERIALS AND METHODS

The implemented classification approach consists of three main consecutive stages. Firstly, the dataset is extracted by using image processing algorithms then data preprocessing procedures are applied to the dataset. Finally, machine learning techniques are used for classification.

A. Dataset

The data used in this paper are adopted with an informed consent obtained from all of the included patients. The dataset is composed of combined ultrasound and ultrasound elastography DICOM images labeled by experienced radiologists and collected over the period from Feb. 2017 to Jun. 2018. A total of 82 images were taken for 34 different patients where some patients have multiple lesions while other patients have only one. Also, some ultrasound elastography are obtained using two different matching materials (oil and gel) between the transducer of an ultrasound imaging system and the breast tissue of the patient. The lesions are labelled as 56 malignant lesions and 26 benign lesions. The images have different sizes where the B-mode image is shown on the left in Fig. 1. Whereas the elastography image is superimposed with the B-mode image as shown on the right. Different levels of measured strains are represented in the elastography images using a color map; where blue represents the highest strain (softest), green represents average strain (intermediate) and red the lowest strain (hardest).

B. Image segmentation

There are two region of interests (ROIs) that can be extracted from each image which are the tumor in the B-mode image and the tumor in the elastography image. Therefore,

978-1-7281-1184-1/19/$31.00 ©2019 IEEE

segmentation of the ROI from the images is performed to be able to extract and measure the features of different tumors for diagnosis.

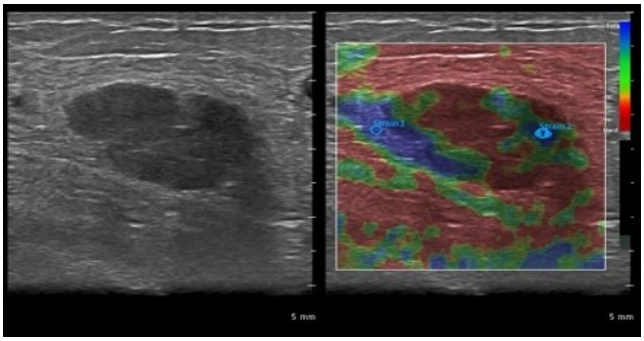

Fig. 1: B-mode image (left) and combined ultrasound elastography image (right)

The images are cropped to separate B-mode images from elastography images. Tumors in B-mode images are segmented using proper threshold for gray levels which separate the tumor from the background. Meanwhile, to extract the tumor in the elastography images, the B-mode image is subtracted from the elastography image followed by masking the elastography image with the ROI in the B-mode. Thereafter, a proper color threshold is applied to extract the region with the lowest strain (hardest). The extracted ROIs and image segmentation are shown in Fig. 2.

C. Feature Extraction

Features are to be extracted from the ROIs to help decide whether the lesion is malignant or benign. A total of 33 features are calculated for both B-mode and elastography images. The extracted features are composed of the geometrical and texture characteristics of either ROIs, their differences or relative values. Some features are related to the texture of the ROI such as mean and standard deviation or geometry of the tumor such as area, perimeter and width-to-height ratio or the quality of images such as contrast to noise ratio and signal to noise ratio. The features are obtained as follows:

- Contrast to Noise Ratio (CNR): It is a measure of the quality of the image. As shown in Eqn. 1, This quantity describes the contrast characteristics between tumor and background [7]:

$$CNR_e = \frac{2(s_1 - s_2)^2}{\sigma_{s_1}^2 + \sigma_{s_2}^2} \quad (1)$$

Where s_1 and s_2 are the average strains of the lesion and background respectively, σ_1 and σ_2 are the standard deviation of tumor and background respectively.

- Signal to Noise Ratio (SNR): Eqn. 2 characterizes the noise between the tumor and the background [7].

$$SNR = \frac{Mean\ strain\ of\ ROI\ (\mu)}{standard\ deviation\ of\ background\ (\sigma)} \quad (2)$$

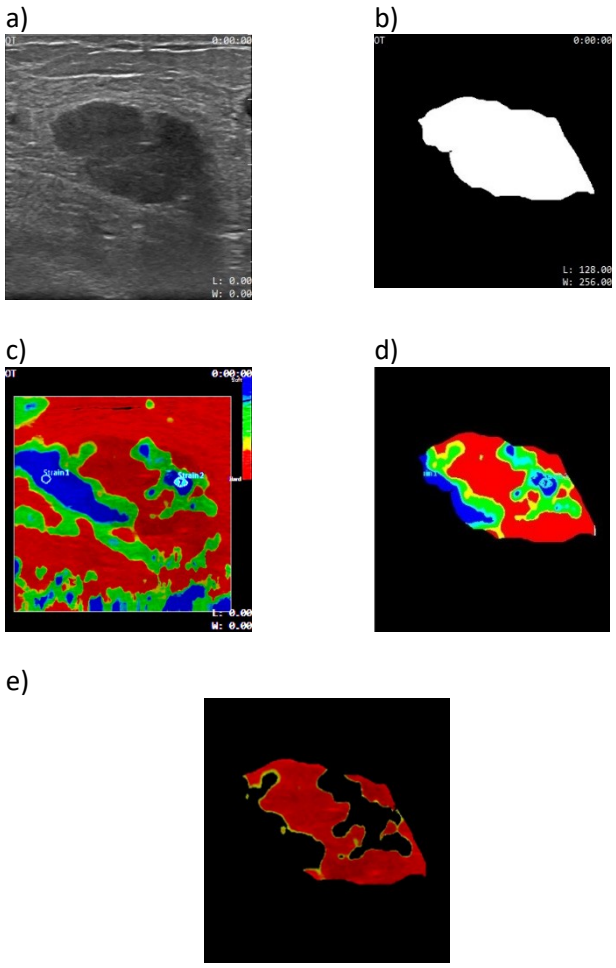

Fig. 1: (a) B-mode with tumor, (b) segmented tumor, (c) elastography image after subtracting the B-mode image from the background, (d) result of masking the elastography image with B-mode ROI and (e) extracted elastography ROI.

- Width to Height Ratio: It is the ratio between the width (W) as the minor axis and the height (H) as the major axis. This ratio is calculated for the elastogram and the B-mode then the difference is calculated as in Eqn. 3.

$$D = \left|\left(\frac{W}{H}\right)_B - \left(\frac{W}{H}\right)_E\right| \quad (3)$$

- Area difference: As shown in Eqn. 4, the area difference is represented by the difference in number of pixels for the tumor in both elastography and b-mode images.

$$A = |N_B - N_E| \quad (4)$$

- Perimeter difference: The perimeter difference represents the difference in length of contour of the tumor in both elastography and b-mode images.
- Solidity: Tumor shape can differentiate between benign and malignant tumors as benign lesions have regular

shapes while malignant lesions have irregular shapes in elastograms.

- Compactness: It is the measure of compression of the tumor in a certain area and is defined by the ratio of the perimeter square to area.

D. Data Preprocessing

After obtaining the resulting dataset from feature extraction. The data is preprocessed by using PCA for dimensionality reduction. This resulted in selecting 18 dimension which are considered the most informative dimensions. The dataset is shuffled and then divided into two main parts where 80% is used as a training set (67 samples) and 20% (17 samples) is used as test set. The training set is used for the model fitting and hyperparameters tuning while the test set is used as a held-out (unseen test set) to evaluate the performance of the model on unseen data.

E. Machine Learning

Applying Machine learning algorithms is considered a crucial step to classify between malignant cases and benign cases with higher accuracy and reliability. For that purpose, different machine learning algorithms are applied. Whereas the classifier which provides the highest possible separability is selected.

Support vector machine (SVM) [9] is selected as the classification algorithm. SVM is firstly applied on the training set for fitting and hyperparameters tuning. Afterwards the resulting model with the selected hyperparameters from the training stage is then tested against the held-out test set for performance evaluation. SVM with radial basis function showed highest accuracy of 94.12%. The performance of this classification scheme in our work compared with other similar work is as shown in Table 1.

Table 1. Accuracy comparison between proposed work and previous work

Reference	Classifier	Accuracy
Proposed work	SVM	94.12%
[10]	Artificial neural networks	90.6%
[11]	SVM	92.3%
[12]	Deep learning	93.4%

III. PERFORMANCE METRICS

The performance metrics that measure how well the model works, consist of three metrics. The first metric is the accuracy; that measures the ratio of predicted labels to the true labels. The second metric is the confusion matrix that measures number of each class that predicted true or confused with other class. Whereas the last metric is the logistic loss which penalizes the classification error.

A. Accuracy

The Main parameter used to measure the performance of the classifiers is the accuracy. This parameter is calculated in terms of true positive (TP), false positive (FP), true negative (TN), false negative (FN) obtained from the confusion matrix of the classifiers. The accuracy [10] can be expressed as shown in Eqn. 5

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

B. Confusion Matrix

The size of confusion matrix is squared, it depends on the number of classes (i.e. for N classes the size of confusion matrix N×N). It shows the number of correctly and incorrectly classified samples.

C. Logistic Loss

The logistic loss function is as defined in Eqn. 6 where f represents the hypothesis function and L is the loss function. Whereas y is the true label [8].

$$L(y, f(x)) = log\left(1 + exp(-yf(x))\right) \quad (6)$$

IV. SIMULATION RESULTS

The confusion matrix shows the predicted labels compared to the actual labels where the benign cases are denoted as 0 and the malignant cases are denoted as 1.

As shown in Fig. 3, the principle diagonal of the confusion matrix represents the correctly classified samples (TP=5 and TN=11) while the secondary diagonal shows the misclassified samples (FP=1 and FN=0).

Fig. 3: Confusion matrix of breast cancer diagnosis (benign and malignant classes)

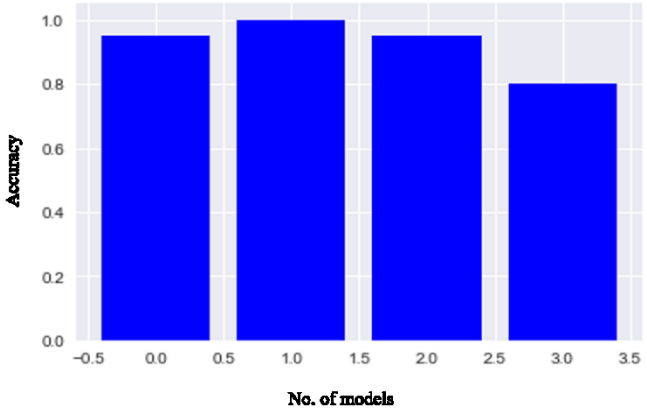

Fig. 4: Accuracy for 4-folds models

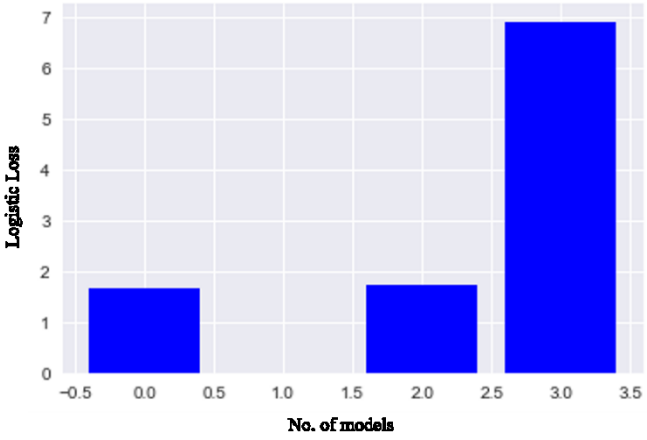

Fig. 5: Logistic loss function for 4-fold

As shown in Fig. 4, the accuracy for 4-folds with average 94.1% at the model number 0, 2 and 100% at model number 1 and for model number 3 the accuracy is the lowest. Logistic loss is the second performance metric for validation of the model error and it is used for optimizing the model using gradient descent algorithm. 4-folds are also used for calculating the loss function as shown in Fig. 5 with an average error of 2.5. From the K-fold validation, the model number "0" and model number "2" have the same accuracy and logistic loss (94.1% and 2 respectively), while the model number "1" and model number "3" aren't more generic than model number "0" and "2", so that they are selected as the fit model. The generated confusion matrix from these models are shown Fig 3.

V. CONCLUSION

In this paper, it is clear that SVM is very helpful for making a decision in breast cancer diagnosis by building a generic and robust model to differentiate between benign and malignant cases. The model has accuracy of 94.1% which is high compared with the literature and logistic loss of 2.5; which means the average error produced by model decisions is 2.5 every 17 samples.

ACKNOWLEDGMENT

This research was partially funded by ONE Lab at Cairo University and Zewail City of Science and Technology.

REFERENCES

[1] Howlader N., and et al., "Breast Cancer Facts and Figures 2017-2018," *American Cancer Society*. Available online at: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

[2] Akay, Mehmet Fatih. "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert systems with applications* 36.2 (2009): 3240-3247.

[3] W. K. Moon, and et al., "Computer-aided tumor diagnosis using shear wave breast elastography," in *Journal of Ultrasonics*, Vol. 78, pp. 125-133, 2017.

[4] H. Zhi, and et al., "Comparison of Ultrasound Elastography, Mammography, and Sonography in the Diagnosis of Solid Breast Lesions," in *Journal of ultrasound in medicine*, Vol. 6, 2007.

[5] Kourou, Konstantina, et al. "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal* 13 (2015): 8-17.

[6] Ahmad, L. Gh, et al. "Using three machine learning techniques for predicting breast cancer recurrence," *J Health Med Inform*4.124 (2013): 3.

[7] R. Jemila Rose, and et al, "Computerized Cancer Detection and Classification Using Ultrasound Images: A Survey," in *International Journal of Engineering Research and Development*,Vol. 5, Issue 7, pp. 36-47, 2013.

[8] James, Witten, Hastie, and Tibshirani, "Statistical Learning," in An Introduction to Statistical Learning, vol. 103, 2013.

[9] Huang, Min-Wei, and et al. "SVM and SVM ensembles in breast cancer prediction," in *PloS one*, Vol. 12, Issue 1, 2017.

[10] Moon, Woo Kyung, and et al. "Analysis of elastographic and B-mode features at sonoelastography for breast tumor classification," in *Ultrasound in medicine & biology*, Vol. 35, Issue 11, pp. 1794-1802, 2009.

[11] Moon, Woo Kyung, and et al. "Computer-aided tumor diagnosis using shear wave breast elastography," in *Ultrasonics 78*, pp. 125-133, 2017.

[12] Moon, Woo Kyung, and et al. "Classification of breast tumors using elastographic and B-mode features: comparison of automatic selection of representative slice and physician-selected slice of images," in *Ultrasound in medicine & biology*, Vol. 39, Issue, 7, pp. 1147-1157, 2013.