Optimal breast cancer diagnostic strategy using combined ultrasound and diffuse optical tomography

K. M. SHIHAB UDDIN,1 MENGHAO ZHANG,2 MARK ANASTASIO,1,3 AND QUING ZHU1,*

1Biomedical Engineering Department, Washington University in St. Louis, 1 Brooking Dr, St. Louis, MO 63130, USA
2Electrical and System Engineering Department, Washington University in St. Louis, 1 Brooking Dr, St. Louis, MO 63130, USA
3Department of Bioengineering, University of Illinois at Urbana-Champaign, 1406 W Green St, Urbana, IL 61801, USA
*zhu.q@wustl.edu

Abstract: Ultrasound (US)-guided near-infrared diffuse optical tomography (DOT) has demonstrated great potential as an adjunct breast cancer diagnosis tool to US imaging alone, especially in reducing unnecessary benign biopsies. However, DOT data processing and image reconstruction speeds remain slow compared to the real-time speed of US. Real-time or near real-time diagnosis with DOT is an important step toward the clinical translation of US-guided DOT. Here, to address this important need, we present a two-stage diagnostic strategy that is both computationally efficient and accurate. In the first stage, benign lesions are identified in near real-time by use of a random forest classifier acting on the DOT measurements and the radiologists’ US diagnostic scores. Any lesions that cannot be reliably classified by the random forest classifier will be passed on to the second stage which begins with image reconstruction. Functional information from the reconstructed hemoglobin concentrations is employed by a Support Vector Machine (SVM) classifier for diagnosis at the end of the second stage. This two-step classification approach which combines both perturbation data and functional features, results in improved classification, as denoted by the receiver operating characteristic (ROC) curve. Using this two-step approach, the area under the ROC curve (AUC) is 0.937 ± 0.009, with a sensitivity of 91.4% and specificity of 85.7%. In comparison, using functional features and US score yields an AUC of 0.892 ± 0.027, with a sensitivity of 90.2% and specificity of 74.5%. Most notably, the specificity is increased by more than 10% due to the implementation of the random forest classifier.

© 2020 Optical Society of America under the terms of the OSA Open Access Publishing Agreement

1. Introduction

Breast cancer is the most common cancer in women worldwide, with approximately 1.67 million new cases each year [1]. Despite improvements in detection and diagnosis, more than half a million patients die from this disease annually. Breast cancer is a spectrum of diseases with different histologic subtypes, grades, and biologic and metabolic activities, resulting in a wide range of functional differences [2]. Benign breast disease also encompasses a heterogeneous group of diseases that vary in vascular content, proliferative index, metabolic activity, and risk of breast cancer [3].

Multiple imaging modalities are currently used for breast cancer screening and diagnosis. X-ray mammography is the predominant modality for both screening and diagnostic imaging. Breast ultrasound (US) is the second most common diagnostic imaging modality and is also used for screening average to moderate risk women with dense breast composition [4–5]. Due to its high cost and limited access, MRI is reserved for screening high risk women and has application to a
very narrow group of diagnostic indications. While the characteristics of malignant and benign breast lesions are well established by conventional imaging techniques [6–7], their overlapping appearances result in approximately one million image-guided breast biopsies each year in the United States, most yielding benign results [8]. An optical tomography system that reveals functional differences in breast abnormalities could greatly improve diagnostic accuracy and reduce the number of benign biopsies.

In the last 20 years, optical breast imaging using diffused light has been widely explored to develop non-invasive imaging tools to detect and diagnose breast cancer, and to predict and monitor its treatment response [9–22]. Initially these systems were investigated as primary or ‘stand-alone’ modalities [5–13]. However, it became clear that the accuracy of DOT could be enhanced for lesion localization and quantification by use of a priori information from other conventional breast imaging modalities, such as mammography/tomosynthesis [19,20], ultrasound (US) [14,16,21] and MRI [17,18]. Dual-modality characterization, incorporating structure from conventional imaging and function from enhanced optical imaging, provides complementary information to improve diagnosis. In particular, ultrasound-guided DOT has demonstrated its translational potential for distinguishing breast cancers from benign lesions [14,21].

One major challenge for dual-modality breast cancer diagnosis is DOT’s relatively slow data processing and image reconstruction speed as compared to the real-time imaging capabilities of US. Near real-time diagnosis is critical for the clinical translation of a US-guided DOT dual-modality technique. The approach described here employs a random forest classifier, an ensemble learning method that has been used widely in medical imaging applications [22]. It makes a decision based on the majority vote of many individual decision trees that are trained on predictive features [23]. Random forest classifiers have demonstrated promising results for computer aided breast cancer diagnosis utilizing US [24], mammogram [25] and biopsy data [26].

In this study, we investigate a two-stage diagnostic strategy for clinically managing breast cancer. The first stage seeks to identify benign lesions in near real-time, based on radiologists’ US scores and DOT measurements in the form of perturbation data that have not undergone image reconstruction. This is accomplished by use of a random forest classifier. Intermediate lesions that cannot be identified as benign with high confidence are flagged, and functional images are subsequently reconstructed off-line from the DOT measurements. In the second stage of the diagnostic strategy, features are extracted from the reconstructed DOT images, and a Support Vector Machine (SVM) classifier is employed for diagnosis. This proposed diagnostic strategy has showed significant improvement over DOT functional feature and US based diagnosis alone by increasing the AUC (area under the ROC curve) from 0.892 to 0.937. To the best of our knowledge, this is the first time a two-step automated diagnostic strategy has been proposed with near real-time assessment capability for the majority of benign lesions.

2. Methods

2.1. Patients and ultrasound BI-RADS grading

A total of 188 patients were studied to evaluate the proposed diagnostic scheme: based on biopsy results, 47 patients had malignant lesions (mean age 59 years; range 34-94 years) and 141 had benign lesions (mean age, 48 years; range 17-82 years). For benign lesions, 32% were fibroadenomas, 24% fibrocystic changes, 9% fat necrosis/inflammatory changes, 14% proliferative lesions, 17% complex cysts, 4% lymph nodes and breast tissue. For malignant lesions, 26% were stage 2 to 4 cancers, and 6% DCIS and 68% stage 1 cancers. The clinical study was approved by the local Institutional Review Board and was compliant with the Health Insurance Portability and Accountability Act. Informed consent was obtained from each patient. Data used in this study were obtained from an earlier study, and data from patients were de-identified [21].
For each lesion, a sequence of US images was obtained and retrospectively reviewed by two radiologists who were blind to the optical results and final diagnosis. The lesions were graded using the Breast Imaging Reporting and Data System (BI-RADS) for US. For each lesion, one of four grades were given, 4A, 4B, 4C and 5 based on the suspicion level of malignancy. BI-RADS 4A refers to ≤10% likelihood of malignancy while 4B, 4C and 5 denotes 10% to 50%, 50% to 95% and ≥95% likelihood of malignancy [6]. In the classification process, all BI-RADS grades (4A to 5) were encoded into numbers from 0 to 3 based on increasing suspicion level, i.e., 0 for 4A, 1 for 4B, 2 for 4C, 3 for 5. These numerical scores from two radiologists were used in the random forest classifier as two additional features to the 12 perturbation features introduced later.

2.2. DOT system

The US guided frequency domain DOT system consisted of a hand-held probe with 9 optical sources and 10 parallel detectors with a US transducer located at the middle [27]. Four laser diodes with wavelengths 740, 780, 808 and 830 nm were sequentially switched by 4×1 and 1×9 optical switches to deliver light modulated at 140 MHz to 9 different source positions on the probe. Backscattered light was collected via 10 light guides on the probe. The output of each detector channel was further amplified and sampled by an analog-to-digital converter (ADC) and stored in a PC. Frequency and phase measurements were extracted from the detected signal using Hilbert transformation. Each data acquisition took 2-3 seconds. Multiple data sets were acquired at the lesion area and the contralateral normal breast at the mirror position, referred to as the reference breast. The entire data acquisition took ~5 minutes to complete, and the perturbation data were calculated immediately after.

2.3. DOT measurements and perturbation features

The DOT perturbation, \( U_{sc} \), is defined as the normalized difference between the lesion and reference measurements, which is related to differential absorption of the lesion and reference normal tissue. For the \( i^{th} \) source-detector pair,

\[
U_{sc}(i) = \frac{A_r(i)e^{i\varphi_r(i)} - A_l(i)e^{i\varphi_l(i)}}{A_r(i)e^{i\varphi_r(i)}} = \frac{A_l(i)e^{i\varphi_l(i)}}{A_r(i)e^{i\varphi_r(i)}} - 1 \quad i = 1, 2, \ldots, m,
\]

where \( m \) is the number of measurements, and \( U_l(i) = A_l(i)e^{i\varphi_l(i)} \) and \( U_r(i) = A_r(i)e^{i\varphi_r(i)} \) are the lesion and reference measurements, respectively. A two-dimensional representation of DOT perturbation measurements is shown in Fig. 1(a) for a benign lesion and Fig. 1(b) for a malignant lesion. The unit circle represents the expected boundary for perturbation data. A convex hull or envelope of the data distribution is marked by a black polygon. For benign lesions, perturbation is skewed towards positive real axis or evenly distributed around both the positive and negative real axes while perturbation for a malignant lesion is skewed toward negative real axis due to high absorption of cancer, which leads to lower ratio of \( \frac{r_A(i)}{r_R(i)} \) (Eq. (1) [28]). This difference in data distribution is quantified by data features extracted from the perturbation that are useful in differentiation of benign and malignant lesions.

Two sets of DOT data features were extracted from the perturbation measurements: morphological features from the convex hull of the data distribution, and histogram features. Four features extracted from the convex hull are: the area, perimeter, moment of inertia and, centroidal polar moments. The moment of inertia, \( I_m \), is a quantitative measurement of the resistance of an object against angular acceleration. The centroidal polar moment, \( I_p \), denotes resistance of the object against torsion or twisting. The definitions of the moment of inertia and centroidal polar moment are as follows:

\[
I_m = \int r^2 dm, \quad I_p = \int r^2 dA,
\]

where \( dm \) and \( dA \) are the differential mass and area elements, respectively, and \( r \) is the distance from the axis of rotation to these elements. Additional details are provided in Appendix A.
For each lesion, all measurements were compiled to generate two separate univariate histograms of the real and imaginary perturbations. A representative example of a univariate histogram for a benign lesion is shown in Fig. 2(a) for real perturbation and Fig. 2(b) for imaginary perturbation. A histogram of a malignant lesion is shown in Fig. 2(d) for real perturbation and 2(e) for imaginary perturbation. From each histogram, six features — the mean, standard deviation, skewness, kurtosis, energy, and entropy — were extracted. In total we obtained 12 features from these two univariate histograms. The real and imaginary perturbations were used together to obtain a bivariate histogram, as shown in Fig. 2(c) for a benign case and 2(f) for a malignant case. Four features — the mean distance from the centroid, standard deviation of distance from the centroid, multivariate skewness, and multivariate kurtosis — were calculated from each bivariate histogram. Two tailed t-test was performed for each feature to calculate p-value, which is an estimate of the predictive capability of the respective feature. All features were ranked in the ascending order of p-values and features with p-value less than 0.05 are used in the classification. A total of 12 features were found significant and used in the random forest classifier described below.

2.4. Random forest classifier

A random forest is an ensemble of decision tree classifiers where each decision tree independently casts a vote for a certain class based on a randomly chosen subset of all features. The final outcome of the forest is based on the majority vote of all the trees. In this study, a total of 14 features, including 12 perturbation features and 2 sets of US BI-RADS scores from two radiologists, were used for classification. The random forest classifier employed in this study consisted of 50 classification and regression trees (CART). Each tree works on 6 randomly selected features out of 14 features. Information gain is used to calculate the best split at each decision tree node. Decision trees can safely handle correlated features too, since once a feature is used to split the samples, information gain on the split samples in the child node will be lower for correlated features [29–30]. Another feature of random forest classifier is that they can possess attractive bias-variance trade-offs if suitably defined. To realize this, in this study, we limited each individual decision tree depth to five and set the number of minimum required samples to split a node to four. To simplify the random forest classifier, Bonferroni correction can be applied to the group of features extracted from the same data set, for example, features from histograms, to obtain corrected p values and reevaluate the significance of each feature. However, this approach may increase the false negative rates on feature selection and miss important features [31]. For decision trees in a random forest classifier, an optimal feature is selected at each node based on information gain between that node and its child nodes. If a feature is not significant, it will not
be selected. A Python open source machine learning library, scikit-learn, was used to build and train random forest classifier.

2.5. Image reconstruction and functional feature extraction

In DOT image reconstruction, the 3-dimensional breast volume to be reconstructed, underneath the 10 cm probe, is represented by voxels, with finer voxels within a lesion area identified by the co-registered US image and coarse voxels in the background region [32]. Fitted optical properties from the contralateral reference breast measurements are used to calculate a weight matrix $W$ (Eqn. (3)) for the voxels. The total absorption of each voxel is reconstructed and then divided by the voxel volume to obtain the differential optical absorption coefficient, $\delta \mu_a$. The inverse problem is linearized by use of the Born approximation to obtain a linear equation relating the changes of the optical absorption coefficients to the perturbation measurements, $U_{sc}$:

$$[U_{sc}]_{m \times 1} = [W_L, W_B]_{m \times n} \begin{bmatrix} \delta \mu_{aL} \\ \delta \mu_{aB} \end{bmatrix}_{n \times 1},$$

$$U_{sc} = WX, W = [W_L, W_B],$$

$$X = \begin{bmatrix} \delta \mu_{aL} \\ \delta \mu_{aB} \end{bmatrix}_{n \times 1},$$

where $W_L$ and $W_B$ are the voxel weights in the lesion and background, respectively; $\delta \mu_{aL}$ and $\delta \mu_{aB}$ are unknown optical properties of voxels in the lesion and background, respectively; and $n$ is the total number of voxels to be reconstructed. The optical absorption coefficients were reconstructed by solving a L2 regularized unconstrained optimization problem using the conjugate gradient method [33].

$$\hat{X} = \arg \min_X \left( \|U_{sc} - WX\|^2 + \frac{\lambda}{2}\|X\|^2 \right),$$

Figure 2. Example histograms from a benign lesion perturbation. (a) univariate histogram for real perturbation, (b) univariate histogram for imaginary perturbation, and (c) bivariate histogram. Example histogram from a malignant lesion perturbation. (d) univariate histogram for real perturbation, (e) univariate histogram for imaginary perturbation, and (f) bivariate histogram.
where, \( \lambda \) is the regularization parameter that is proportional to tumor size and largest singular value of the weight matrix, \( W \).

Oxy- and deoxy- hemoglobin concentrations, \((C_{HbO_2}, C_{Hb})\), were calculated from four wavelength absorption maps using the value of the extinction coefficient, \( \varepsilon \), for different wavelengths,

\[
\begin{bmatrix}
\mu^740_a \\
\mu^780_a \\
\mu^808_a \\
\mu^830_a
\end{bmatrix}
= 
\begin{bmatrix}
\varepsilon^740_{Hb} & \varepsilon^740_{HbO_2} \\
\varepsilon^780_{Hb} & \varepsilon^780_{HbO_2} \\
\varepsilon^808_{Hb} & \varepsilon^808_{HbO_2} \\
\varepsilon^830_{Hb} & \varepsilon^830_{HbO_2}
\end{bmatrix}
\begin{bmatrix}
C_{Hb} \\
C_{HbO_2}
\end{bmatrix}.
\] (5)

Functional features were extracted from the reconstructed hemoglobin maps. Three features were calculated from all lesion images, \( C_{HbO_2} \), \( C_{Hb} \), and total hemoglobin \( C_{tHb} \). Since \( C_{tHb} \) is the summation of \( C_{HbO_2} \) and \( C_{Hb} \), there are only two independent features. In general, \( C_{Hb} \) is much lower than \( C_{HbO_2} \), for breast lesions, thus it is less robust in computation. Therefore, we have chosen \( C_{tHb} \) and \( C_{HbO_2} \) for classification.

The light shadowing effect was also used as a functional imaging feature [34]. In reflection geometry, large tumors are more likely to show light shadow effect in reconstructed DOT images because photons are absorbed more by top portion and less photons penetrate deeper. This results in the reduced signal to noise ratio received at longer source and detector pairs and therefore low reconstructed absorption values at deeper layers. To quantify the light shadow effect, the shadow parameter was calculated as the ratio of average \( C_{tHb} \) calculated from the topmost layer in depth and the average of the underlying layers. Examples of a malignant lesion (a)-(b) and a benign lesion (c)-(d) are given in Fig. 3 to demonstrate the light shadow effect. The shadow parameter or ratio of the malignant lesion is 4.52, and the ratio of the benign lesion is 2.12. The three functional features, \( C_{tHb} \), \( C_{HbO_2} \) and light shadow parameter, are the functional features used in the second step of diagnosis by the SVM classifier.

### 2.6. Support vector machine

The Support Vector Machine (SVM) is a widely used binary classifier that finds the optimum hyperplane to separate two classes by maximizing the margin of error. In this study, we use a linear SVM. For a collection of feature vectors, \( \{x_i\} \) and associated class labels, \( y_i \in \{-1, +1\} \), we find the optimal hyperplane \( w^T x + b = 0 \), where, \( w \) is the weight of support vectors and \( b \) is the bias term. The following optimization problem is solved to find the weight, \( w \),

\[
\min_w C \sum_{i=1}^{n} \max[1 - y_i(w^T x + b), 0] + ||w||_2^2,
\] (6)

where, \( C \) is the regularization term. A Python open source machine learning library, scikit-learn, was used to build and train SVM classifier.

### 2.7. Two-step classification

Breast lesions were diagnosed in two steps. Immediately after data acquisition, perturbation features were extracted and US BI-RADS scores were obtained from the radiologists. These perturbation features and BI-RADS scores were used in a random forest classifier to identify lesions having a high probability of being benign. The total number of decision tree votes required to declare a benign lesion was set very high so that the false negative rate would be very small or nonexistent in near real-time assessment. Two-thirds of the malignant samples and the same number of benign samples were used for training, and the rest were used for testing. Recall that our sample set contained 47 malignant cases and 141 benign cases. The training set comprised 32 malignant and 32 benign cases. The test set was comprised of the remaining 15
Fig. 3. Light shadowing effect observed in large tumors. An example of a malignant lesion (a)-(b), and an example of a benign lesion (c-d). (a) co-registered US image, (b) reconstructed total hemoglobin map of a large malignant lesion. The shadow parameter or ratio of the average topmost layer total hemoglobin and the average of underlying layers is 4.52. (c) co-registered US image, and (d) reconstructed total hemoglobin map of a benign lesion. The ratio of the average topmost layer total hemoglobin and the average of underlying layers is 2.12. In hemoglobin map, each 2D slice is spatial x-y image of 4.5 cm × 4.5 cm and slices 1 to 7 are at 0.5 cm to 3.5 cm depth, in 0.5 cm increment.

malignant cases and 109 benign cases. In this first step, all 32 benign and 32 malignant cases were used for training. Hyperparameter tuning was performed by 5-fold cross-validation on the training set.

Image reconstruction and functional feature extraction were done for lesions with intermediate diagnoses. Based on functional features, these samples were classified using a SVM classifier. Here, all 32 malignant cases were again used in training; however, for benign cases, lesions with intermediate malignancy probability, not filtered in first step, were used in training. Again, hyperparameters were selected by 5-fold cross-validation performance. The test set of 15 malignant and 109 benign cases was not employed for training or validation. Thus, the test data were unseen by both the random forest and SVM classifiers. This entire two-step process was repeated 20 times for different random train-test split as illustrated in Fig. 4. Note that, the chart presented in Fig. 4 is the workflow that we reanalyze previously collected data in this study. In the first step in training, perturbation features and US BI-RADS scores of all training samples are used to train a random forest classifier. Each decision tree outputs a binary decision of either benign or malignant for each training sample. In general, if more than half of the decision trees in the forest provide a benign decision, that sample is assumed to be benign. However, in this classification scheme, the threshold for the total number of decision tree votes to determine benignity was set as high as possible to avoid false negatives in the first step. A greedy search was applied to find the threshold. Initially the threshold, i.e., the number of votes required to
To determine benignity, was set to the maximum number of decision trees, which is 50. Then the threshold was decreased in steps of 1 as long as there were no false negatives.

**Fig. 4.** Two-step diagnosis scheme, with each step bounded by dashed rectangles.

Using this approach, the minimum threshold that provided 100% training sensitivity was achieved in the first step. In contrast, in testing, a sample was classified as ‘confirmed benign’ in first step only when the total number of trees voting ‘benign’ was greater than or equal to the threshold. In the second step of diagnosis, image reconstruction was done to obtain hemoglobin maps for the remaining samples. The maximum total hemoglobin, maximum oxy-hemoglobin and light shadow parameters were extracted from the maps. These functional features were then used to classify rest of the samples using the SVM classifier.

The objective function for random forest is misclassification cost. If true label is $y_i$ and predicted label is $y'_i$, for $i$ - $\text{th}$ sample, the objective function is,

$$
\arg\min_{x_i, \text{split}} \sum_{i=1}^{n} f(y_i, y'_i), \quad \text{where}, \quad f(y_i, y'_i) = \begin{cases} 
0, & \text{if } y_i = y'_i \\
w_{y_i, y'_i}, & \text{otherwise}
\end{cases},
$$

(7)

Here, $n$ is the number of samples, $x_i, \text{split}$ is the split value to be used for splitting the parent node for feature $x_i$, and $w_{y_i, y'_i}$ is the misclassification cost of a sample originally with label $y_i$ predicted as $y'_i$. Detailed explanation of the misclassification cost and minimization of the cost function can be found in [35]. For second step using SVM, the objective function is given in Eq. (6). In the first step, random forest classifier filters out benign cases with low malignancy probabilities. In the second step, the SVM classifies the remaining cases. The trade-off between these two steps are controlled by the threshold of decision tree votes to determine benign in first step. If the threshold is high, only few benign lesions can be classified and the false negative rate in first step is lower. If the threshold is low, more benign lesions can be identified but the false negative rate can be higher. Please note that the threshold for total number of decision trees is not fixed. For each random train-test split, the threshold was based on minimizing false negatives in the training data.
Finally, to evaluate and visualize the importance of each feature at both steps of diagnosis, we have used random forest to calculate feature importance for all features including perturbation features, functional features and US BI-RADS ranks. In the random forest, the importance of each feature is the average information gain due to that feature only across all the trees [36]. First 15 most important features found through this method are visualized in Fig. 5. The importance of these features is normalized to the summation, thus they add up to 1. In terms of computation speed, for 20 runs of random train-test split, random forest classification takes approximately 5 to 6 seconds, while SVM takes only 1 to 1.5 seconds. The computation is performed in an Intel core i5 3.0 GHz CPU with 8 GB RAM.

![Feature Importance](image)

Fig. 5. Importance of top 15 features

### 2.8. Performance evaluation

To evaluate the performance of the classification algorithms, for each sample in the test set, we computed the probability of malignancy from the respective classifier. The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) was used as the performance measure to evaluate the classifiers. Specifically, the threshold for determining malignant or benign class based on malignancy probability is varied to produce ROC curves. For classifications using SVM, the probability is obtained based on the distance of a test sample from the optimal hyperplane. The distance is passed as an argument to a sigmoid function to obtain a probability estimate. Then we can vary the probability threshold from 0 to 1 to generate true positive and false positive rates and obtain a ROC curve. For random forest, probability of prediction is related to the proportion of decision trees that voted for that predicted label. In the two-step approach, the first step yields true negatives and false negatives with zero true positives and false positives. In the second step, true positives and false negatives generated from second step as well as summation of true negatives and false negatives from both steps are used to generate a combined ROC.

Twenty runs with different random train-test splits were performed for each classifier. The mean AUC denoted how well the classifier could separate benign and malignant classes and the standard deviation indicated the robustness of the classifier for varying training and testing data sets. Sensitivity and specificity were calculated at a threshold of 0.5 from the mean ROC curve.
To evaluate the radiologists’ performance, the sensitivity and specificity were calculated based on BI-RADS scores: 4A and 4B were grouped as benign, and 4C and 5 as malignant.

3. Results

3.1. Perturbation feature selection and random forest classifier

A total of 20 features were extracted from perturbation data and box plots and p-values of all features are shown in Fig. 6. The first 12 features are statistically significant and have used for the first step random forest classifier. 64.8% (±4.7%) of the benign lesions were identified by the random forest classifier with 1.9% false negative rate in testing, which was 70 cases on average out of 109 benign cases. Note that low malignancy lesions were filtered out and diagnosed as benign, thus there was no true positive or false positive rate.

3.2. Clinical study results

The BI-RADS for both radiologists are shown in Table 1. Using BI-RADS scores only and grouping 4A and 4B as benign, and 4C and 5 as malignant, the sensitivities for radiologist I and II were 70.9% (±0.3%) and 85.6% (±0.2%), and the specificities were 90.8% (±2.2%) and 63.5% (±2.4%), respectively. The ROC curves for radiologist I and II are shown in Fig. 7(a) and 7(b), with AUC value 0.848 ± 0.003 and 0.783 ± 0.031 respectively. The blue curve and the light blue shade denote the mean and standard deviation of 20 ROC curves obtained from 20 runs.

| Table 1. The number of lesions for each category of BI-RADS for both radiologists |
|-----------------------------------|---|---|---|---|
|                                  | 4A | 4B | 4C | 5  |
| Radiologist I                    |    |    |    |    |
| Benign                           | 29 | 99 | 10 | 3  |
| Malignant                        |  0 | 13 |  8 | 26 |
| Radiologist II                   |    |    |    |    |
| Benign                           | 22 | 92 | 20 |  7 |
| Malignant                        |  1 | 11 | 16 | 19 |

Using functional features only in the SVM classifier, the AUC was 0.781 ± 0.048, as shown in Fig. 7(c) with a sensitivity of 82.5% (±4.2%) and specificity of 72.9% (±1.0%). Using BI-RADS scores along with functional features in the SVM classifier improved the AUC to 0.892 ± 0.027 (Fig. 7(d)), with a sensitivity of 90.2% (±1.9%) and specificity of 74.5% (±1.3%).

The proposed two-step diagnosis significantly improved the AUC to 0.937 ± 0.009 (Fig. 7(e)), with a sensitivity of 91.4% (±0.6%) and specificity of 85.7% (±0.8%). In the first step of the two-step method, 64.8% benign samples were classified as benign by the random forest classifier. Even though a zero false negative rate was enforced in training, 1.9% (±0.6%) of malignant samples were misclassified as benign in testing in the first step.

It is interesting to compare the performance of the aforementioned ROC curves with the ROC curve of random forest with all features included in one step. As shown in Fig. 7(f), the AUC was 0.923 ± 0.020, which is statistically the same as the two-step classification (see 3.3). The sensitivity and specificity were 90.3 ± 0.7% and 82.1 ± 1.2%, respectively. However, the proposed two-step approach provides near real-time diagnosis capability. AUCs for all different diagnostic schemes are summarized in Table 2.

3.3. Comparison of ROC curves

To evaluate the performance of different diagnostic methods, we used DeLong’s method to compare different ROC curves obtained from different diagnostic methods. An open source MATLAB software package written by Sun et al. is used [36]. The p-values obtained from the method are given in Table 3. As seen from the table, the performance of functional feature with
US BI-RADS scores is statistically significantly better than that with functional feature only. The proposed two-step reconstruction also performs significantly better than functional feature with US BI-RADS. The proposed two-step diagnostic method is statistically the same as using all the features combined, however, the advantage of the two-step method is the near real-time diagnosis and immediate recommendation for patients who have benign lesions.
Fig. 7. ROC curves of different classification methods. (a) BI-RADS score for radiologist I, (b) BI-RADS score for radiologist II, (c) functional feature only, using SVM. (d) BI-RADS score with functional features using SVM. (e) Proposed two-step diagnostic scheme. (f) Random forest with all features in one-step.

|                          | Functional Feature only | Functional Feature with US BI-RADS | Two-Step Diagnostic Scheme | Random forest with all features in one-step |
|--------------------------|-------------------------|-----------------------------------|---------------------------|--------------------------------------------|
| Functional Feature only  | —                       | —                                 | P<0.001                   | P<0.001                                   |
| Functional Feature with US BI-RADS | P<0.001                | —                                 | P<0.001                   | P=0.002                                   |
| Two-Step Diagnostic Scheme | —                       | —                                 | P=0.267                   | P=0.002                                   |

4. Discussion and summary

In summary, a novel breast cancer diagnostic strategy based on a two-step classification strategy was proposed and validated with a large pool of patient data. This strategy involves near real-time automated assessment using a random forest classifier to filter out highly probable benign lesions
based on perturbation data and US BI-RADS scores. Lesions that cannot be identified as benign with high confidence are flagged, and their functional images are subsequently reconstructed off-line from the corresponding DOT measurements. In the second stage of the diagnostic strategy, features are extracted from the reconstructed DOT images, and a Support Vector Machine (SVM) classifier is employed for diagnosis. Functional feature extraction can take up to two hours including manual US image segmentation and optical image reconstruction with artifact evaluation. However, these steps are critical to provide high diagnostic accuracy.

The random forest classifier reliably predicted more than half of benign lesions in near real time, shortly after perturbation features were extracted and radiologist’s BI-RADS scores were available. In practice, BI-RADS are typically available within a few minutes after the patient exam. Such rapid diagnosis helps advance clinical management by identifying highly probable benign lesions and allowing the physicians to comfortably recommend follow-up instead of biopsy or surgical removal of the lesions. Additionally, US BI-RADS evaluation is highly dependent on the radiologist’s experience; while the random forest classifier combines sensitive perturbation data with the BI-RADS to provide an improved diagnosis over that of a radiologist alone.

The two-step diagnosis scheme improves the specificity of a breast cancer diagnosis over a diagnosis based on the BI-RADS score and DOT-derived functional parameters only. This improvement is due to the diagnosis of highly probable benign lesions by the random forest classifier. A lower standard deviation across multiple cross-validations indicated this approach is very robust to different training and testing datasets and hence more reliable. Introducing perturbation features in the first step improved the overall diagnostic performance and facilitated better clinical management of the benign lesions to reduce unnecessary biopsies. Although a hemoglobin map is reconstructed from perturbation data, the tumor size and location provided by co-registered US and the breast tissue optical background properties are also used in the reconstruction process. The tumor size and location define the fine mesh area and location, and the background optical properties are used to calculate the weight matrix. Thus, for similar perturbation data, the reconstructed functional features can be different for different background properties, and lesion dimensions and locations. Our results suggest that this additional information employed when reconstructing functional features is valuable to further differentiate benign and malignant lesions.

For large benign lesions, even if the absorption coefficient is high, the hemoglobin concentration map shows less light shadowing and a more uniform distribution in depth, which is critical in differentiating large benign lesions and malignant tumors. For low grade carcinomas (14.63% in this study), the detection sensitivity of DOT can be lower due to the low level of tumor angiogenesis, however, the distorted tumor morphology evaluated by US BI-RADS is very helpful in improving diagnosis. Additionally, certain types of fibroadenomas are vascularized and present as false positives to DOT, however, the fibroadenomas’ well circumscribed morphology in US image can help rule out malignance.

This study has the limitation that radiologists’ evaluations were done on stationary ultrasound images. Real time assessment of ultrasound images while examining the patient may improve the overall diagnostic performance. Additionally, with other diagnostic information, such as mammograms and patient family history, the overall diagnostic performance can be further improved. In current practice, suspicious lesions found from x-ray mammograms are referred to ultrasound for targeted examination of the lesion regions. Based on real-time ultrasound, the attending radiologist will provide BI-RADS score and make a recommendation of follow-up or biopsy immediately after the US exam. Our proposed study flow will immediately provide recommendation based on the optimal strategy reported in this manuscript which can minimize the biopsy recommendations for a large portion of benign lesions. This is a direction that we are pursuing in on-going clinical studies.
The majority of the biopsy patient population has benign findings and the benign to malignant ratio is 3:1 in this study. We have chosen balanced data sets in training because a bias in training set can increase the false negative rate. With a bias in training set, we could have a classifier biased toward the majority class, which is benign findings, and then we would have more false negatives in the testing result. We did majority under sampling [37] to reduce the training bias to ensure an unbiased classifier. The only problem of bias in testing set is that accuracy of the classifier will be skewed toward the majority class. Thus, we didn’t use accuracy as performance evaluation but AUCs for the proposed algorithms.

The 1.9% false negative rate of the first step random forest classifier warrants discussion. In the twenty runs of random train-test splits, two malignant cases were categorized as benign in some tests. One case is a small ductal carcinoma (5 mm measured by US) with low vascular content and another is a small mixed dual and lobular carcinoma (3 mm measured by US) with low vascular content. Both lesions are intermediate grade and two radiologists scored them as 4B. This type of false negative is difficult to avoid, however, including x-ray mammogram reading could add another parameter to identify this type of small tumors. Additionally, 3 to 6 month follow up recommendation would also allow a small window to monitor the development of this type of small tumors.

The proposed novel two-step diagnostic strategy employing a random forest classifier as a first step to filter out low suspicious benign lesions during patients’ US exam has great potential to streamline breast diagnostic work flows by suggesting short-term follow-ups rather than biopsy. Based on a large patient pool, 64.8% of the benign lesions were identified by the first step random forest classifier with 1.9% false negative rate. The next step using an SVM classifier combining DOT total hemoglobin functional maps with other diagnostic image features, provides high overall performance, AUC of 0.937, in breast cancer diagnosis. The reported two-step diagnostic strategy can be generalized to other modality guided diffused optical tomography for the optimal management of breast cancer diagnosis.

Appendix A

Equation (2) is a general definition of moment of inertia and centroidal polar moment. Exact methods of calculating moments for a 2D polygon is described in [38]. If the polygon has n vertices, \((x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\), then, moment of inertia, 
\[
I_m = \frac{1}{24} \sum_{i=1}^{n} (x_i y_{i+1} + y_i x_{i+1} + x_{i+1} y_i + 2x_i y_i + x_{i+1} + y_{i+1}) * 
\left(\sum_{i=1}^{n} (x_i * y_{i+1} - y_i * x_{i+1})\right)
\]

centroidal polar moment,
\[
I_p = I_{xx} + I_{yy} - \frac{1}{2} (x c^2 + y c^2) \left(\sum_{i=1}^{n} (x_i y_{i+1} - y_i x_{i+1})\right)
\]

Where, \(xc\) and \(yc\) are x and y coordinates of the centroid, and \(I_{xx}, I_{yy}\) are defined as follows,
\[
I_{xx} = \frac{1}{12} \sum_{i=1}^{n} (x_i^2 + y_i x_{i+1} + x_{i+1} y_i + 2x_i y_i + x_{i+1} + y_{i+1}) * (x_i y_{i+1} - y_i x_{i+1})
\]
\[
I_{yy} = \frac{1}{12} \sum_{i=1}^{n} (y_i^2 + y_i y_{i+1} + y_{i+1} x_i + 2y_i x_i + y_{i+1} + x_{i+1}) * (x_i y_{i+1} - y_i x_{i+1})
\]

Funding

National Cancer Institute (R01CA228047, R01EB002136).

Acknowledgments

The authors acknowledge and are grateful for funding support for this work from the NIH (R01EB002136, R01 CA228047). The authors also thank James Ballard for proof reading of the manuscript.
Disclosures

The authors declare that there are no conflicts of interest related to this article.

References

1. National Breast Cancer Foundation, 2019, https://www.nationalbreastcancer.org/
2. K. Polyak, “Heterogeneity in breast cancer,” J. Clin. Invest. 121(10), 3786–3788 (2011).
3. M. Guray and A. Sahin, “Benign Breast Diseases: Classification, Diagnosis, and Management,” Oncologist 11(5), 435–449 (2006).
4. R. I. Hooley, L. M. Scoutt, and L. E. Philpotts, “Breast Ultrasonography: State of the Art,” Radiology 268(3), 642–659 (2013).
5. J. Okello, H. Kisenbo, S. Bugeza, and M. Galukande, “Breast cancer detection using sonography in women with mammographically dense breasts,” BMC Med. Imaging 14(1), 41 (2014).
6. American College of Radiology. ACR BI-RADS® Atlas Fifth Edition Quick Reference: “Ultrasound, Mammography, Magnetic Resonance Imaging, BI-RADS® Assessment Categories,” https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BI-RADS-Reference-Card.pdf (Accessed on November 13, 2019).
7. A. Goel and F. Gaillard, “Benign and malignant characteristics of breast lesions at ultrasound,” Mayo Clin. Proc. 85(3), 274–279 (2010).
8. C. I. Flowers, C. O’Donoghue, D. Moore, A. Goss, D. Kim, J. H. Kim, S. G. Elias, J. Fridland, and L. J. Esserman, “Reducing false-positive biopsies: a pilot study to reduce benign biopsy rates for BI-RADS 4A/B assessments through testing risk stratification and new thresholds for intervention,” Breast Cancer Res. Treat. 139(3), 769–777 (2013).
9. B. J. Tromberg, A. Cerussi, N. Shah, M. Compton, A. Durkin, D. Hsiang, J. Butler, and R. Mehta, “Imaging in breast cancer: diffuse optics in breast cancer: detecting tumors in pre-menopausal women and monitoring neoadjuvant chemotherapy,” Breast Cancer Res. 7(6), 279 (2005).
10. R. Choe, S. D. Konecky, A. Corlu, K. Lee, T. Durduran, D. R. Busch, S. Pathak, B. J. Czernecki, J. Tchou, D. L. Fraker, A. Demichele, B. Chance, S. R. Arridge, M. Schweiger, J. P. Culver, M. D. Schnall, M. E. Punt, M. A. Rosen, and A. G. Yodh, “Differentiation of benign and malignant breast tumors by in-vivo three-dimensional parallel-plate diffuse optical tomography,” J. Biomed. Opt. 14(2), 024020 (2009).
11. S. P. Poplack, T. D. Tosteson, W. A. Wells, B. W. Pogue, P. M. Meaney, A. Hartov, C. A. Kogel, S. K. Soho, J. J. Gibson, and K. D. Paulsen, “Electromagnetic Breast Imaging: Results of a Pilot Study in Women with Abnormal Mammograms,” Radiology 243(2), 350–359 (2007).
12. X. Intes, “Time-domain optical mammography SoftScan: initial results,” Acad Radiol. 12(8), 934–947 (2005).
13. L. Spinelli, A. Torricelli, A. Pifferi, P. Taroni, G. Danesini, and R. Cueddu, “Characterization of female breast lesions from multi-wavelength time-resolved optical mammography,” Phys. Med. Biol. 50(11), 2489–2502 (2005).
14. Q. Zhu, P. U. Hegde, A. Ricci, M. Kane, E. B. Cronin, Y. Ardestehpour, C. Xu, A. Aguirre, S. H. Kurtzman, P. J. Deckers, and S. H. Tannenbaum, “Early-stage invasive breast cancers: potential role of optical tomography with US localization in assisting diagnosis,” Radiology 256(2), 367–378 (2010).
15. S. Ueda, N. Nakamiya, K. Matsura, T. Shigekawa, H. Sano, E. Hirokawa, H. Shimada, H. Suzuki, M. Oda, Y. Yamashita, O. Kishino, I. Kuji, A. Osaki, and T. Saeki, “Optical imaging of tumor vascularity associated with proliferation and glucose metabolism in early breast cancer: clinical application of total hemoglobin measurements in the breast,” BMC Cancer 13(1), 514 (2013).
16. J. S. Choi, M. J. Kim, J. H. Youk, J. H. Moon, H. J. Suh, and E. K. Kim, “US-guided optical tomography: correlation with clinicopathologic variables in breast cancer,” Ultrasound Med Biol. 39(2), 233–240 (2013).
17. V. Ntziachristos, A. G. Yodh, A. G. Yodh, M. D. Schnall, and B. Chance, “MRI-guided diffuse optical spectroscopy of malignant and benign breast lesions,” Neoplasia 4(4), 347–354 (2002).
18. M. A. Mastanduno, J. Xu, F. El-Ghussein, S. Jiang, H. Yin, Y. Zhao, K. E. Michaelisen, K. Wang, F. Ren, B. W. Pogue, and K. D. Paulsen, “Sensitivity of MRI-guided near-infrared spectroscopy clinical breast exam data and its impact on diagnostic performance,” Biomed. Opt. Express 5(9), 3103 (2014).
19. Q. Fang, J. Seh, S. A. Carp, G. Boverman, E. L. Miller, D. H. Brooks, R. H. Moore, D. B. Kopans, and D. A. Boas, “Combined optical and X-ray tomosynthesis breast imaging,” Radiology 258(1), 89–97 (2011).
20. V. Krishnaswamy, K. E. Michaelisen, B. W. Pogue, S. P. Poplack, I. Shaw, K. Defrietias, K. Brooks, and K. D. Paulsen, “A digital x-ray tomosynthesis coupled near infrared spectral tomography system for dual-modality breast imaging,” Opt. Express 20(17), 19125 (2012).
21. Q. Zhu, A. Ricci, P. Hegde, M. Kane, E. Cronin, A. Merkulov, Y. Xu, B. Tavakoli, and S. Tannenbaum, “Assessment of Functional Differences in Malignant and Benign Breast Lesions and Improvement of Diagnostic Accuracy by Using US-guided Diffuse Optical Tomography in Conjunction with Conventional US,” Radiology 280(2), 387–397 (2016).
22. M. Hsuni, A. Ittisam, A. Idri, J. M. C. de Gea, and J. L. F. Alemán, “Reviewing ensemble classification methods in breast cancer,” Computer Methods and Programs in Biomedicine 177, 89–112 (2019).
23. L. Breiman, “Random forests,” Machine learning 45(1), 5–32 (2001).
24. 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 Med Biol. 42(4), 980–988 (2016).
25. J. Liu, J. Chen, X. Liu, and J. Tang, “An investigate of mass diagnosis in mammogram with random forest,” Fourth International Workshop on Advanced Computational Intelligence, IEEE, 638–641 (2011).

26. F. K. Ahmad and N. Yusoff, “Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier,” in 13th International Conference on Intelligent Systems Design and Applications, IEEE, 121–125 (2013).

27. Vavadi, A. Mostafa, F. Zhou, K. S. Uddin, M. Althobaiti, C. Xu, R. Bansal, F. Ademuyiwa, S. Poplack, and Q. Zhu, “Compact ultrasound-guided diffuse optical tomography system for breast cancer imaging,” J. Biomed. Opt. 24(2), 1 (2018).

28. K. S. Uddin and Q. Zhu, “Reducing image artifact in diffuse optical tomography by iterative perturbation correction based on multiwavelength measurements,” J. Biomed. Opt. 24(5), 1 (2019).

29. A. Liaw and M. Wiener, “Classification and regression by random Forest,” R news. 2(3), 18–22 (2002).

30. C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. Marquèz, B. Gruber, B. Lafourcade, P. J. Leitão, and T. Münkemüller, “Collinearity: a review of methods to deal with it and a simulation study evaluating their performance,” Ecography 36(1), 27–46 (2013).

31. K. Rothman, “No adjustments are needed for multiple comparisons,” Epidemiology 1(1), 43–46 (1990).

32. Q. Zhu, N. Chen, and S. H. Kurtzman, “Imaging tumor angiogenesis by use of combined near-infrared diffusive light and ultrasound,” Opt. Lett. 28(5), 337 (2003).

33. K. S. Uddin, A. Mostafa, M. Anastasio, and Q. Zhu, “Two step imaging reconstruction using truncated pseudoinverse as a preliminary estimate in ultrasound guided diffuse optical tomography,” Biomed. Opt. Express 8(12), 5437 (2017).

34. C. Xu and Q. Zhu, “Light shadowing effect of large breast lesions imaged by optical tomography in reflection geometry,” J. Biomed. Opt. 15(3), 036003 (2010).

35. J. Schiffrs, “A classification approach incorporating misclassification costs,” IDA 1(1), 59–68 (1997).

36. Y. Saeys, T. Abeel, and Y. Van de Peer, “Robust feature selection using ensemble feature selection techniques,” Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, 313–325 (2008).

37. X. Y. Liu, J. Wu, and Z. H. Zhou, “Exploratory undersampling for class-imbalance learning,” IEEE Transactions on Systems, Man, and Cybernetics 39(2), 539–550 (2009).

38. H. J. Sommer III, “POLYGEOM: Geometry of a Planar Polygon,” (2005).