Computer-Aided Theragnosis Based on Tumor Volumetric Information in Breast Cancer

Mehrdad J. Gangeh, *Senior Member, IEEE,* Simon Liu, Hadi Tadayyon, and Gregory J. Czarnota

***Abstract*— *Objective:* A computer-assisted technology has recently been proposed for the assessment of therapeutic responses to neoadjuvant chemotherapy in patients with locally advanced breast cancer (LABC). The system, however, extracted features from individual scans in a tumor irrespective of its relation to the other scans of the same patient, ignoring the volumetric information. This study addresses this problem by introducing a novel engineered texton-based method in order to account for volumetric information in the design of textural descriptors to represent tumor scans. *Methods:* A noninvasive computer-aided-theragnosis (CAT) system was developed by employing multiparametric QUS spectral and backscatter coefficient maps. The proceeding was composed of two subdictionaries: one built on the "pretreatment" and another on "week *N*" scans, where *N* was 1, 4, or 8. The learned dictionary of each patient was subsequently used to compute the model (histogram of textons) for each scan of the patient. Advanced machine learning techniques including a kernel-based dissimilarity measure to estimate the distances between "pretreatment" and "mid-treatment" scans as an indication of treatment effectiveness, learning from imbalanced data, and supervised learning were subsequently employed on the texton-based features. *Results:* The performance of the CAT system was tested using statistical tests of significance and leave-one-subject-out (LOSO) classification on 56 LABC patients. The proposed texton-based CAT system indicated significant differences in changes between the responding and nonresponding patient populations and achieved high accuracy, sensitivity, and specificity in discriminating between the two patient groups early after the start of treatment, i.e., on weeks 1 and 4 of several months of treatment. Specifically, the CAT system achieved the area under curve of 0.81, 0.83, and 0.85 on weeks 1, 4, and 8, respectively. *Conclusion:* The proposed texton-based CAT system accounted for the volumetric information in "pretreatment" and "mid-treatment" scans of each patient. It was demonstrated that this attribute of the CAT system could boost its performance compared to the cases that the features were extracted from solely individual scans.**

Manuscript received April 13, 2017; accepted April 27, 2018. Date of publication May 23, 2018; date of current version July 30, 2018. This work was supported in part by the Natural Science and Engineering Research Council of Canada under Postdoctoral Fellowship PDF-454649-2014, in part by the Terry Fox Foundation, and in part by the Canadian Institutes of Health Research. *(Corresponding author: Mehrdad J. Gangeh.)*

M. J. Gangeh is with the Global Innovation Team, Ernst & Young (EY), Palo Alto, CA 94301 USA (e-mail: mehrdad.gangeh@utoronto.ca).

S. Liu and H. Tadayyon are with the Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 2M9, Canada, and also with the Department of Physical Sciences, Sunnybrook Health Sciences Center, Toronto, ON M4N 3M5, Canada (e-mail: simon.liu@sunnybrook.ca; hadi.tadayyon@sunnybrook.ca).

G. J. Czarnota is with the Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 2M9, Canada, with the Department of Radiation Oncology, University of Toronto, Toronto, ON, M5G 2M9 Canada, with the Department of Radiation Oncology, Sunnybrook Health Sciences Center, Toronto, ON M4N 3M5, Canada, and also with the Department of Imaging Research—Physical Sciences, Sunnybrook Health Sciences Center, Toronto, ON M4N 3M5, Canada (e-mail: gregory.czarnota@sunnybrook.ca).

Digital Object Identifier 10.1109/TUFFC.2018.2839714

***Index Terms*— Computer-aided theragnosis (CAT), locally advanced breast cancer (LABC), neoadjuvant chemotherapy, personalized medicine, quantitative ultrasound (QUS), textons, treatment response monitoring.**

I. Introduction

RESPONSE of cancer therapy depends on many factors that are specific to individual patients being treated. Most of these factors center around responses to a treatment regimen based on the patient's genetic profile. Other factors such as patient's age and environment, tumor grade, and stage also play important roles. Instead of following a single regimen for all patients at the same stage of a specific cancer, personalized medicine aims to tailor the most effective treatment regimen as early as possible to patient-specific factors.

Locally advanced breast cancer (LABC) is a large category of breast cancer in which patients have the following: tumor size greater than 5 cm with positive axillary lymph nodes, and/or bulky, fixed axillary adenopathy, and may involve breast skin and/or chest wall [1]. Given that LABC is often considered inoperable, neoadjuvant (preoperation) chemotherapy is usually administered as the start to possible breast conserving cancer treatment, a technique which was pioneered in the setting of LABC [2], [3]. Since outcomes in terms of patient survival are linked to tumor response to chemotherapy, it is vital to periodically monitor the chemotherapy response and make adjustments as quickly as possible.

Ideally, cancer treatment response monitoring at the cellular level is based on imaging detectable structures inside the tumors, consequently allowing the monitoring of changes in tumor microstructure as cell death occurs. A variety of modalities presently under investigation, namely, positron emission tomography (PET) [4], magnetic resonance imaging (MRI) [5], diffuse optical imaging (DOI) [6], and ultrasound use the manifestations of physics phenomena that can probe tumor structure and size. Although these methods can probe tumor physiology, only size is used clinically to assess response. Ultrasound is one of the most effective modalities in terms of cost, imaging time, and portability. In comparison, MRI and PET require large capital investment and injection of exogenous agents, with the latter exposing the patient to potentially high dose of ionizing radiation. Lower resolution DOI has yet to gain clinical utilization.

The plausibility of using conventional-frequency ultrasound for response monitoring is due to fairly recent developments in quantitative ultrasound (QUS) methods. These methods aim to capture acoustic scattering derived tumor

properties, termed QUS biomarkers, which detect changes in tumor microstructure early on during a course of treatment (days as opposed to months as carried out in clinical practice using standard imaging methods) [7]–[12]. Early QUS imaging [13] studies used high frequency ultrasound (>20 MHz) to obtain images of tumor response, but imaging depth limitations have led to the rapid exploration of QUS tumor monitoring using conventional mid- to low-frequency ultrasound (1–20 MHz) [8], [10], [12]. In QUS, the raw radio frequency (RF) data is used instead of ultrasound B-mode images, since the latter are image-processed and log-compressed data sets that are unreliable in depicting all frequency-dependant microstructure changes in the tumor during cell death [7]. QUS methods are used to analyze the entire frequency-dependent power spectrum, capturing the effects of scattering microstructures more effectively than B-mode intensity images. Even though B-mode imaging is a useful convention for the visual observation of artifacts inside patients and other abnormalities, QUS importantly provides ultrasound machine-independent measurements of tumor characteristics.

A. Computer-Aided Theragnosis

Computer-aided theragnosis (CAT) is an emerging technology to assist radiation oncologists to monitor and assess therapeutic cancer responses early after the start of treatment administration [14]. This computer-assisted system would enable to noninvasively classify cancer patients as responders or nonresponders during the course of treatment. Medical images acquired from functional imaging modalities, such as QUS parametric maps, are used as inputs to CAT systems. A CAT system typically needs to perform several tasks including feature extraction from input images, measuring the dissimilarities between the baseline (i.e., "pretreatment") and "mid-treatment" scans, and the classification of patients as responders or nonresponders based on the measured dissimilarities (see Fig. 1). Ground-truth labels are determined based on the ultimate clinical and histological analysis.

Recently, Gangeh *et al.* [14] proposed a CAT system based on QUS methods at conventional frequencies for the assessment of LABC patients' responses to neoadjuvant chemotherapy. However, one major shortcoming of the CAT system developed in [14] is that the features have been extracted from individual scans irrespective of their relations to the other scans of the same patient. However, the hypothesis is that including volumetric information into the design of the feature extractor can potentially improve the performance of the CAT system in classifying patient responses to treatment. In particular, it is important to extract discriminative features from multiple scans of a tumor considering their interrelations, and also their relations to those scans, with which they will eventually be compared, that is the "mid-treatment" scans. On the other hand, it is known that the responses developed in tumors as a result of treatment administration are often heterogeneous [15], which suggests using textural features to describe these responses [10], [14]. Therefore, in this paper, textural features were extracted from the 2-D QUS

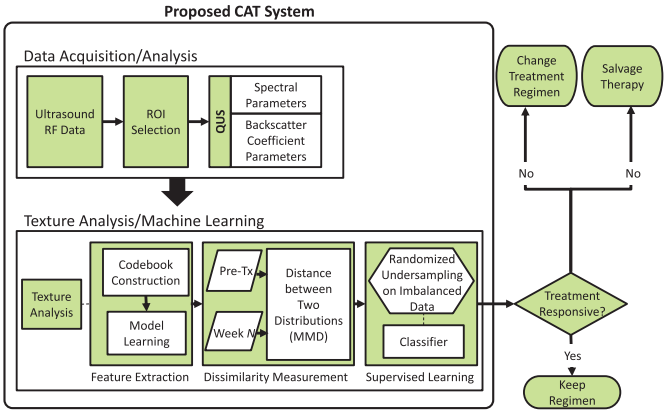

Fig. 1. Schematic of the proposed CAT system used for the classification of LABC patients as responders or nonresponders.

parametric maps. Furthermore, in order to account for volumetric information, a texton-based approach—a data-driven dictionary of primitive elements constructed on "patches" extracted from input images—was adapted. One dictionary of textons was built for each patient using all the "pretreatment" and "mid-treatment" scans of that particular patient. The dictionary, therefore, was representing the volumetric information in all the scans of a patient. Subsequently, model histograms of textons for the scans of the particular patient were derived using the constructed dictionary as detailed in Section II. The texton-based method has led to the state-of-the-art results in texture analysis on benchmark data sets [16] and in other medical applications such as the classification of lung parenchyma in CT images [17], [18]. The investigation here was conducted to demonstrate the high performance of a CAT system using the texton-based method to characterize QUS parametric maps in the classification of tumor responses.

In the developed CAT system, the dissimilarities between "pretreatment" scans as the baseline, and weeks 1, 4, and 8 "mid-treatment" scans were quantified for each patient and subsequently submitted as inputs to a classifier. Differences were calculated using a kernel-based dissimilarity measure called the maximum mean discrepancy (MMD), which was first proposed by Gretton *et al.* [19] to address two-sample problem, i.e., to test whether two distributions p and q are different on the basis of samples drawn from each of them. In the application of cancer response monitoring, p and q are the distributions of the data samples taken from "pretreatment" and "mid-treatment" at a specific time interval after the start of treatment, respectively. Due to the prevalence of responding patients compared to nonresponding ones, a compensation for learning from imbalanced data using random undersampling [20] was also incorporated before the submission of the dissimilarities to a classifier. Fig. 1 illustrates the schematic of the complete CAT system developed in this paper.

B. Contributions

As a summary, the main contributions of this study are as follows:

1) the design and development of a complete noninvasive CAT system based on QUS methods, which accounted for the volumetric information in multiple scans of each patient, by using a data-driven texture method;
2) a novel engineering texton-based method specifically adapted to the application of cancer response monitoring by building one dictionary/codebook per patient, and learning the scan models based on the patient-specific dictionary;
3) quantifying the different ultrasound characteristics of responding tumors compared to nonresponding tumors as early as week 1 in patients with LABC receiving neoadjuvant chemotherapy when using the proposed CAT system.

A preliminary version of this study has been reported as an abstract in [21].

II. Methods

A. Image Analysis Using Texture Features

The viability of using QUS as a tumor response modality is based upon the wealth of frequency information that is contained in the raw RF images. The use of texture analysis on parametric maps directly derived from raw RF signals takes advantage of the texture features found in images derived from all available frequency data. The texton approach [16], [22] is considered a dictionary learning approach where the atoms of the dictionary (textons) are used to define the texture models, by the means of a histogram of textons. In contrast, local binary patterns (LBPs) [23], which are considered as another state-of-the-art method for texture analysis, rely on predefined operators. An advantage of using data-driven texture analysis over those based on parametric responses to predefined operators is their adaptability to the type of data set. In addition, the dictionary-based texton method can compute the basis using all the scans in one patient, thereby incorporating volumetric information into model features as opposed to extracting features from individual scans using texture methods such as the LBPs.

1) Feature Extraction Using Texton Approach: The main concept of textons was originally introduced by Julesz [24] to represent the elements of texture perception. The first proposal for a complete texture classification system using textons was proposed only 20 years later by Leung and Malik [25]. The approach was subsequently further optimized by several researchers as published in [16], [22], and [26], resulting in the state-of-the-art results on benchmark texture data sets as well as medical images [17]. Three major representations associated with the texton approach consist of raw-pixel representation [16], [22], filter banks [25], [26], and the Markov random field [16], where the central pixel is modeled using the neighboring pixels. The steps for the texton approach, irrespective of representation used to describe local image information, consist of learning a dictionary/codebook of textons and the construction of a model for each image using dictionary elements. These steps are described in this section as the core principles to texton dictionary learning approach in the context of cancer response monitoring.

a) Construction of codebook: To construct the codebook, patches of a specified size were extracted from each image in the data set. These patches were converted into an appropriate representation such as raw-pixel representation and vectorized. The vector elements representing the patches were then submitted to a clustering algorithm (e.g., k-means) to cluster all the patch representations into a specified number of dictionary elements (the textons). In the context of cancer response monitoring, for each patient, subdictionaries were learned from the baseline ("pretreatment") scans and "mid-treatment" scans separately, and composed into a single dictionary that was used for model learning for the patient's regions of interest (ROIs) in the next step. The learning of dictionaries for each patient was separated due to the patient-specific nature of ultrasound tumor data. The proposed CAT system represents a response classification system based upon differences between "pretreatment" and "mid-treatment" scan planes. In order to assign a representative measure that would quantify the difference between scans at a specific time during treatment versus the "pretreatment" scans for a patient, the use of only a single patient's ROIs for dictionary learning provides a metric to better represent a dissimilarity feature that was independent of other individuals. If the dictionary was constructed using clustering of all features from all patients, the dissimilarity measure used to represent a patient would not have been the characteristic of solely the individual, potentially leading to a poorer classification performance.

b) Model learning: After constructing a codebook of textons for each patient, the models (feature sets) were built for the ROIs of each patient using the texton codebook constructed in the previous step for that particular patient. To this end, patches of the same size as in the codebook construction step were extracted for analysis by sliding a window pixel-by-pixel over each ROI of the patient. The patches were subsequently converted into the appropriate representation as used in the previous step and vectorized. Eventually, a histogram of textons was computed, as the model for each ROI, by comparing the extracted patches with the codebook textons constructed for each corresponding patient and finding the closest match using a similarity measure, such as Euclidean distance, to update the corresponding bin in the histogram of textons. Therefore, the model for an ROI was a histogram where the bin labels corresponded to textons, and the frequency of patches most similar to each of the textons determined the shape of the histogram. Fig. 2(a) and (b) illustrate the codebook construction and model learning steps in the texton approach.

B. Maximum Mean Discrepancy

Cancer response monitoring involved categorizing the patients into two categories: treatment responders and nonresponders, thus classification was based upon measuring the dissimilarity of each time group after chemotherapy with the "pretreatment" scans. This choice of categorization was the reason why the model histograms were not arbitrarily submitted to train the classifier. Furthermore, a patient being categorized as a responder or nonresponder depends on the changes since starting treatment, thus submitting the

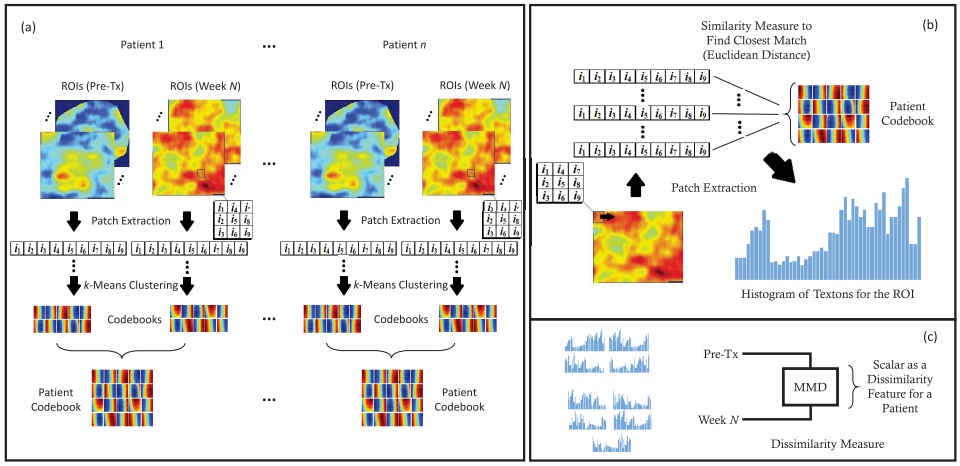

Fig. 2. Outline of the texture analysis/machine learning algorithm using the texton approach. (a) Codebook construction. (b) Model learning for each ROI. (c) Calculating the feature for each patient which represents the dissimilarity between "pretreatment' and "mid-treatment" ROIs.

features of "pretreatment" and "mid-treatment" ROIs separately will not prove useful in quantifying changes within these two populations that is crucial for characterizing treatment efficacy. Assigning a value for each patient that quantifies the differences between the scans taken "pretreatment" and "mid-treatment" is synonymous to identifying the treatment effectiveness, i.e., the larger the distance, the more success in treatment [7], [13]. Finding the optimal dissimilarity measure that most accurately calculates the differences between the two populations is, therefore, important in the design of a CAT system.

One of the simplest and most straightforward methods to measure the distance between the two distributions is to calculate the distance between the cluster means using

$$\mathrm{d}(p,q) = \|\mathbf{E}(p) - \mathbf{E}(q)\|_2^{\frac{1}{2}} \tag{1}$$

where $\mathbf{E}$ is the expectation function, p and q are the two distributions. The main drawback of the metric given in (1), which is equivalent to Euclidean distance or ℓ_2 norm, is that it only takes into account the first order statistics of the data samples taken from p and q. Therefore, if the two distributions have the same mean values, they cannot be discriminated using (1) even if, e.g., their standard deviations (second-order statistics) are different.

One approach to overcome this problem is to first map the data to a higher dimensional feature space and then compute (1) in the augmented feature space. By computing the expectation function in the augmented feature space, higher order statistics of the two distributions can effectively be taken into account resulting in potentially enhanced discrimination. This idea was effectively and efficiently implemented by Gretton *et al.* [19], leading to a nonparametric (i.e., making no assumption on the distributions p and q) kernel-based metric in reproducing kernel Hilbert spaces called MMD.

To provide a formal description, let $X = \{x_i\}_{i=1}^n$ and $Y = \{y_i\}_{i=1}^m$ be data samples drawn independently and identically distributed (i.i.d.) from p and q, respectively. A feature mapping function φ can be defined such that $X \sim p, X \xrightarrow{\varphi} \varphi(X)$, and similarly $Y \sim q, Y \xrightarrow{\varphi} \varphi(Y)$, which maps the data to a high-dimensional feature space. By computing (1) in this space, a metric was computed with the following formulation:

$$\begin{aligned}\mathrm{MMD}(\varphi, p, q) &= \|\mathbf{E}\{\varphi(p)\} - \mathbf{E}\{\varphi(q)\}\|_2^{\frac{1}{2}} \\ &= \left[\mathbf{E}\{[\varphi(X) - \varphi(Y)]^\top [\varphi(X) - \varphi(Y)]\}\right]^{\frac{1}{2}} \\ &= [\mathbf{E}\{\varphi(X)^\top \varphi(X) - 2\varphi(X)^\top \varphi(Y) \\ &\quad + \varphi(Y)^\top \varphi(Y)\}]^{\frac{1}{2}}. \end{aligned} \tag{2}$$

In practice, to compute (2) using a finite number of data samples $X = \{x_i\}_{i=1}^n$ and $Y = \{y_i\}_{i=1}^m$ taken from the two distributions p and q, respectively, the following empirical formulation for MMD can be used:

$$\mathrm{MMD}(\varphi, X, Y) = \left[\frac{1}{n^2}\sum_{i,j} k(x_i, x_j) - \frac{2}{nm}\sum_{i,j} k(x_i, y_j) + \frac{1}{m^2}\sum_{i,j} k(y_i, y_j)\right]^{\frac{1}{2}} \tag{3}$$

where $k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$.

Intuitively, it is expected that empirical MMD will be small when $p = q$ and large when the two distributions are far apart. It can be computed in quadratic time, i.e., for $n + m$ data samples the cost of computation is $\mathcal{O}((n+m)^2)$ time, which is reasonably low for a real-time computation. The empirical MMD was adapted in this study to compute the dissimilarity between the "pretreatment" and "mid-treatment" samples of each patient at a specific time interval after the start of treatment [illustrated in Fig. 2(c)].

C. Learning from Imbalanced Data and Classification

In machine learning, there are situations where the number of data samples is not equally distributed across different classes. For example, most clinical data sets naturally include many more data samples in the healthy class than the cancerous or abnormal class. Likewise, in this study, the

number of responders was about three times more than the nonresponders. Therefore, a simple yet efficient method for learning from imbalanced data was adapted here, which was based on random undersampling. The method used the entire minority class and a random subsample of the majority class with the same size as the minority class for training the classifier. Undersampling was preferred over oversampling as the latter tends to lead to poorer classification on test data as a consequence of overfitting [27], [28].

The diagnosis tool in our study was constructed by training a classifier on balanced data using random undersampling and with the labels of "responder" and "nonresponder" obtained from clinical and histological analysis. The focus of this study was mainly on feature extraction design using data-driven texture methods, and on the dissimilarity measure between the "pretreatment" and "mid-treatment" scans (MMD versus ℓ_2 norm) in the design of the CAT system. Therefore, since among the classifiers tested, a naïve Bayes classifier achieved higher performance compared with a support vector machine (SVM) with radial basis function (RBF) kernel, and a k-NN classifier, the results were only reported for this classifier, in order to enable focusing more on the comparison among the combinations of feature extraction and dissimilarity measures used. The conjuncture on a higher performance of naïve Bayes classifier compared with k-NN or SVM with the RBF kernel is that the two latter classifiers have one (k in k-NN) or two (kernel width γ and tradeoff parameter C in SVM with the RBF kernel) parameters to tune. Considering the small sample size in this paper, this parameter tuning may result in overfitting that may deteriorate the performance of the CAT system.

III. Experimental Setup

The developed CAT system was validated on 56 LABC patients who received neoadjuvant chemotherapy. The patient data were the same as what was reported in [14], and is briefly explained in Section III-A. However, in addition to spectral parametric maps, which are based on linear models on power spectra computed on raw RF data, more complex models named backscatter coefficient (BSC) parametric maps were also included in this study.

A. Patient and Ultrasound Data

1) Patient Data: The study was conducted in accordance with institutional research ethics approval from Sunnybrook Health Sciences Centre. The study involved 56 female patients from the ages of 29 to 67 (49 ± 9/mean $\pm$ SD) years with a diagnosed LABC who participated in the study after written consent. Tumor sizes ranged between 2 and 15 (6.5 ± 3.3) cm. In order to confirm a cancer diagnosis, biopsy was performed on all patients prior to therapy.

Ultrasound B-mode and RF data were acquired using an L14-5/60 transducer with a transmission frequency of 10 MHz and a center frequency of $\sim$7 MHz, connected to a Sonix RP (Ultrasonix, Vancouver, BC, Canada) ultrasound device. The RF sampling rate was set at 40 MHz and the depth of focus for each patient was kept constant throughout the study. Each patient was scanned before treatment and again three other times during chemotherapy, i.e., at weeks 1, 4, and 8. Scans were acquired to cover the entire tumor at approximately 1-cm spacing, resulting in 3–5 image planes across the tumor.

2) Histology: Patient mastectomy samples were obtained after surgery and mounted on whole-mount [29] 5" $\times$ 7" pathology slides with subsequent haematoxylin and eosin (H&E) staining. High-resolution images were obtained using a confocal scanner (TISSUEscope^{TM}, Huron Technologies, Waterloo, ON, Canada) at a resolution of 2 μm.

Ground-truth classification labels of responders and nonresponders were obtained on the basis of the modified response evaluation criteria in solid tumors [14], [30]. Based on these criteria, patients were categorized into two groups: responders ($n_1 = 42$), and nonresponders ($n_2 = 14$). This categorization was performed by assessing the changes in overall tumor size reduction and residual tumor cellularity, which were determined from clinical reports, diagnostic MRI (initial and residual tumor size), and histopathology data (residual tumor size and cellularity). Fig. 3(b) depicts representative histology slides for a responding and a nonresponding patient at low and high magnifications.

3) Quantitative Ultrasound Analysis: As elaborated in [14], the analysis of ultrasound RF data was performed across all scan planes with identifiable tumor regions by selecting an ROI containing the tumor. The analysis included the following steps: 1) computation of fast Fourier transform on Hanning-gated RF segments; 2) normalization of power spectra in order to remove system related variations including system transfer functions and transducer beam forming; and 3) correction of power spectra for frequency-dependent depth-based attenuation [31].

a) Spectral parameters: Linear regression analysis was performed on the normalized and attenuation-corrected power spectrum [32] within a -6 dB bandwidth around the transducer's center frequency (4.5–9.0 MHz). Three spectral parameters obtained from the linear regression analysis were: 1) the regression fit (in dBr) at the center frequency of -6-dB window, termed mid-band fit (MBF); 2) the intercept of the fit at 0 MHz, termed the 0-MHz intercept or spectral intercept; and 3) the slope of the fit, termed spectral slope. These spectral parameters are physically related to ultrasound backscatter power, acoustic scatterer concentration, and effective acoustic scatterer size, respectively, as established by Lizzi *et al.* [32].

b) Backscatter coefficient (BSC) parameters: The same normalized and attenuation-corrected power spectrum was used to estimate the BSC parameters over the ROIs using the reference phantom technique [33]. Subsequently, by a least-squares fitting of the Gaussian form factor to the BSC, parameters of this form factor corresponding to the maximum coefficient of determination, including average acoustic concentration (AAC) and average scatterer diameter (ASD), were computed as outlined by Insana and Hall [34]. The AAC parameter is related to relative acoustic impedance and effective scatterer number density, and the ASD, as the name implies, is related to the scatterer size [34].

Representative "pretreatment" and "mid-treatment" MBF and AAC parametric maps are depicted in Fig. 3(a) over the

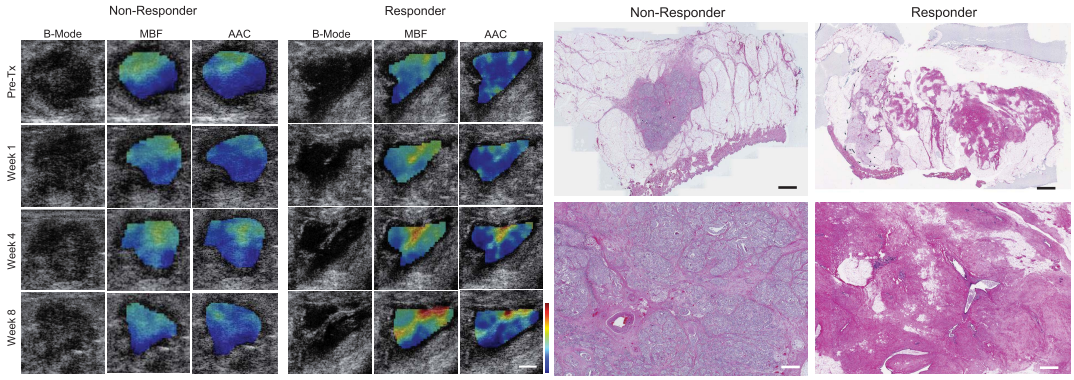

Fig. 3. (a) Representative B-mode images, MBF, and average AAC parametric maps overlaid on corresponding B-mode images for a typical responder and nonresponder LABC patients. The images were acquired from the same nominal tumor locations prior to the start of neoadjuvant chemotherapy (Pre-Tx) and during the course of treatment, i.e., on weeks 1, 4, and 8. The color bar represents an increase of −16 to 28 dBr for MBF and 24 to 86 dB/cm^3 for AAC; the scale bar represents ∼1 cm. (b) Representative high-magnification microscopic images of whole-mount histopathology slides for a responder (left) and a nonresponder patient. The scale bar represents 250 μm.

course of treatment for typical responding and nonresponding patients.

B. Alternative Features and Dissimilarities

The performance of the proposed CAT system using the texton method for feature extraction and the MMD metric as a dissimilarity measure was compared with two other texture feature representations: LBPs [23] and gray level co-occurrence matrices (GLCM) [35], and with the mean of intensity ("MeanInt") as features for parametric ultrasound maps. For the texton, LBP, and GLCM features, the MMD and ℓ_2 norm were compared as distance measures. The difference of magnitude, denoted by Δ, was used as a dissimilarity measure associated with the "MeanInt" feature representation.

C. Implementation Details

Experiments were performed on all spectral and BSC parameters. However, here the results were reported for the MBF as a representative spectral parameter and the AAC as a representative BSC parametric map. These two parameters have shown better performance in two pilot studies on limited LABC data [36], [37] using "MeanInt" or GLCM as features, the results which were also confirmed in this study on a larger data set and using more sophisticated texture and machine learning methods employed in the proposed CAT system.

As described in Section II-A1, codebook construction and model learning for the texton-based approach was performed per patient, meaning that one codebook was constructed for each patient separately. This was done by composing two subdictionaries learned separately on "pretreatment" and "week N" scans. To build each subdictionary, 500 patches were randomly extracted from each scan in "pretreatment" or "week N" of a patient. Raw-pixel representation was directly used without applying any filters, as recommended by [16], and odd patch sizes from 3×3 to 15×15 pixels and five values of k for k-means clustering, i.e., $k = \{10, 20, \ldots, 50\}$ were used. One set of MMD values were computed for each pair of k and patch size values, resulting in 35 sets [5 (different k values) $\times$ 7 (different texton sizes)] of MMD values. A feature selection using sequential forward selection (SFS) algorithm in a wrapper framework [38] was used on the training set to find the best k and patch size combination.

Several settings of the GLCM can be controlled including the relative distance (d) and orientation (θ) of pixel pairs to be compared [35], and the number of gray levels in scaling. In our setup, the GLCMs were constructed from a combination of four orientations: 0°, 45°, 90°, and 135° and distances from 1 to 10 pixels, using eight levels of gray scaling, and pairing order invariance. Second order textural statistics were extracted from each GLCM, i.e., the contrast, correlation, energy, and homogeneity and a corresponding vector of four scalars was obtained for each pair of (d, θ), which was subsequently averaged over all 40 (10 distances $\times$ 4 orientations) values, and considered as the feature representation for each ROI. This is the same setup as being used in [36] on the LABC data.

As for the LBP texture method, rotation invariant uniform two LBPs ($\text{LBP}^{\text{riu2}}_{P,R}$) [23] were used ($P$ is the number of neighboring pixels and R is the radius of a circular local binary operator) at three different (P, R) pair values: (8,1), (16,2), and (24,3) to realize a multiresolution implementation of the LBP. The histograms generated as the models/features for the parametric maps were of $P + 2$ bins, i.e., 10, 18, and 24 at the three different resolutions, respectively. The same feature selection as the one used for the texton-based method was employed on the training set to find the best resolution.

An RBF kernel was selected for the computation of empirical MMD given in (3). To avoid the overfitting due to tuning using small data samples, a self-tuning approach similar to what has been explained in [39] was used to set the kernel width for the RBF kernel. That is, the median of all (Euclidean) distances among the "pretreatment"and "mid-treatment" samples was considered as the kernel width for each subject at a specific treatment interval.

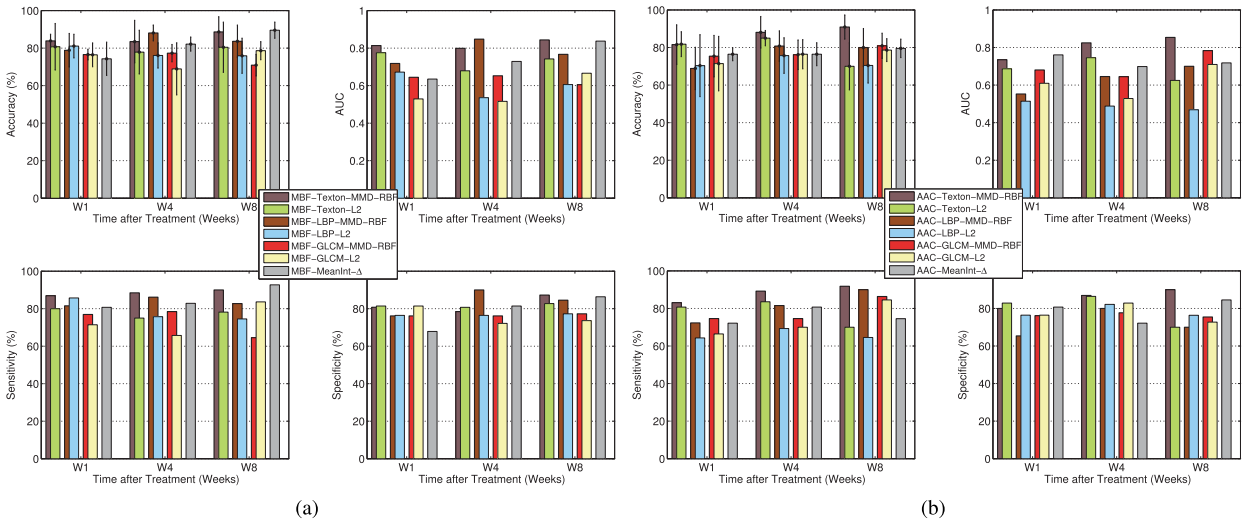

Fig. 4. Classifier accuracy, AUC (the ROC), sensitivity (related to the percentage of nonresponders correctly identified), and specificity (related to the percentage of responders correctly identified) for binary classification between responders/nonresponders over three different time periods during neoadjuvant chemotherapy administration for (a) MBF as a representative spectral, and (b) average AAC as a representative BSC parametric map. Leave-one-out at subject level was used; results shown are averaged over ten runs over different random undersampled sets of the majority class (responders). Error bars represent the standard deviations over ten runs.

For different combinations of parametric maps, feature sets, and dissimilarity measures, the following naming convention is used here:

[Parametric Map]-[Feature]-[Distance Measure]-[Kernel]

e.g., MBF-LBP-MMD-RBF.

IV. Results

A. Classification of Therapeutic Response

The classification of patients' responses to treatment at each time interval during the course of treatment (weeks, 1, 4, and 8) was performed by the submission of computed dissimilarities to a naïve Bayes classifier. The classification was performed in an LOSO scheme. All required tunings as well as feature selections were performed on the training set in each fold of LOSO based on validation classification accuracy. Since the majority class was undersampled during the training of the classifiers to compensate for the imbalanced distribution of the data samples in the two classes of responders and nonresponders, the experiments were repeated 10 times using different subsampled sets of the majority class, and the results were averaged and reported.

Four measures were used in order to evaluate the performance of the classifier: accuracy, area under curve (AUC) of the receiver operating characteristic (ROC), sensitivity, and specificity, which are reported in Fig. 4 and Tables I and II. The measure of focus is the sensitivity, which measures the true positive rate of the classifier corresponding to the correctly identified nonresponders. Identifying refractory patients correctly and switching to more effective treatment regimens is necessary for preventing more serious cancer effects like metastasis. In contrast, if a responder is incorrectly identified as a nonresponder, additional assessment for determining the patient's status will not prove harmful. Fig. 4 summarizes classification results for the MBF and AAC using all feature-distance combinations tested in this study for all time groups, with results for ℓ_2 norm overlaid over their MMD counterparts. As the treatment duration lengthens, predictions for the ultimate responsiveness of the treatment improve, at least for the well performing Texton-MMD approach. The LBP-MMD approach, which was the only other well performing method in this study, followed the same trend, except for the MBF parameter, it instead peaked in performance at week 4. Another rival method, which obtained promising results, was the "MeanInt" feature for MBF, which was inferior to the other methods prior to week 8, but demonstrated the highest performance once that time was reached.

The proposed texton method using MMD as a kernel-based distance measure achieved the best results for 3 out of 4 of

TABLE I

Classification Results for Prediction of Treatment Response for All Patients at Week 1. Results Are Listed From the Highest Classification Accuracy to the Lowest Accuracy

Method	Accuracy (%)	AUC	Sensitivity (%)	Specificity (%)
MBF-Texton-MMD	83.9	0.814	86.9	80.8
AAC-Texton-ℓ_2	81.8	0.687	80.7	82.9
AAC-Texton-MMD	81.5	0.735	83.1	80.0
MBF-LBP-ℓ_2	81.1	0.672	85.7	76.4
MBF-Texton-ℓ_2	80.7	0.776	80.0	81.4
MBF-LBP-MMD	78.9	0.718	81.5	76.2
MBF-GLCM-MMD	76.5	0.644	76.9	76.2
AAC-MeanInt-Δ	76.4	0.761	72.1	80.7
MBF-GLCM-ℓ_2	76.4	0.529	71.4	81.4
AAC-GLCM-MMD	75.4	0.681	74.6	76.2
MBF-MeanInt-Δ	74.3	0.635	80.7	67.9
AAC-GLCM-ℓ_2	71.4	0.610	66.4	76.4
AAC-LBP-ℓ_2	70.4	0.514	64.3	76.4
AAC-LBP-MMD	68.9	0.553	72.3	65.4

TABLE II

CLASSIFICATION RESULTS FOR PREDICTION OF TREATMENT RESPONSE FOR ALL PATIENTS AT WEEK 4. RESULTS ARE LISTED FROM THE HIGHEST CLASSIFICATION ACCURACY TO THE LOWEST ACCURACY

Method	Accuracy (%)	AUC	Sensitivity (%)	Specificity (%)
AAC-Texton-MMD	88.1	0.825	89.2	86.9
MBF-LBP-MMD	88.1	0.848	86.2	90.0
AAC-Texton-ℓ_2	85.0	0.746	83.6	86.4
MBF-Texton-MMD	83.5	0.799	88.5	78.5
MBF-MeanInt-Δ	82.1	0.729	82.9	81.4
AAC-LBP-MMD	80.8	0.646	81.5	80.0
MBF-Texton-ℓ_2	77.9	0.679	75.0	80.7
MBF-GLCM-MMD	77.3	0.653	78.5	76.2
AAC-MeanInt-Δ	76.4	0.699	80.7	72.1
AAC-GLCM-ℓ_2	76.4	0.529	70.0	82.9
AAC-GLCM-MMD	76.2	0.645	74.6	77.7
MBF-LBP-ℓ_2	76.1	0.536	75.7	76.4
AAC-LBP-ℓ_2	75.7	0.488	69.3	82.1
MBF-GLCM-ℓ_2	68.9	0.516	65.7	72.1

TABLE III

TEST OF STATISTICAL SIGNIFICANCE (UNPAIRED TWO-SAMPLE T-TEST OR TWO-WAY MANN–WHITNEY TEST DEPENDING ON THE NORMALITY OF THE TWO POPULATIONS OF RESPONDERS AND NONRESPONDERS). THE p-VALUES ARE SHOWN FOR RESPONDERS VERSUS NONRESPONDERS DURING THE COURSE OF TREATMENT.
* DENOTES $p < 0.05$; ** DENOTES $p < 0.01$

Param. Map	Feature-Distance	Week 1	Week 4	Week 8
MBF	Texton-MMD-RBF	0.038*	0.023*	0.019*
	LBP-MMD-RBF	0.528	0.011*	0.048*
	GLCM-MMD-RBF	0.199	0.089	0.989
	Texton-ℓ_2	0.289	0.274	0.332
	LBP-ℓ_2	0.947	0.198	0.607
	GLCM-ℓ_2	0.962	0.233	0.354
	MeanInt-Δ	0.551	0.068	0.064
AAC	Texton-MMD-RBF	0.148	0.026*	0.003**
	LBP-MMD-RBF	0.260	0.046*	0.130
	GLCM-MMD-RBF	0.358	0.452	0.073
	Texton-ℓ_2	0.215	0.339	0.324
	LBP-ℓ_2	0.769	0.436	0.589
	GLCM-ℓ_2	0.603	0.643	0.503
	MeanInt-Δ	0.947	0.339	0.130

the measures, including a very high sensitivity value of 87% in week 1 on MBF parametric map. In general, almost all of the best overall results were obtained by the texton-based approach using both ℓ_2 norm and MMD, with the Texton-MMD approach clearly achieving the highest AUC.

Fig. 4 indicates that a dissimilarity measure using a kernel-based metric such as MMD combined with textons outperformed the "MeanInt" method, that is, representing the entire ROI with a single scalar value. Compared to using a simple Euclidean distance measure, MMD achieved significantly better results, except for GLCM using the MBF parameter on week 8. Results favored the analysis of the MBF parameter early in the treatment, at week 1, and the use of the AAC backscatter parameter for subsequent weeks, at least starting from week 4.

B. Statistical Significant Difference Between Responders and Nonresponders

A D'Agostino-Pearson test using GraphPad Prism (GraphPad Software, La Jolla, CA, USA), along with graphical analysis, was conducted as a normality test for dissimilarity values for each class at every time group. Unpaired t-tests for normal distributions and two-way Mann–Whitney tests otherwise, using a significance level of $\alpha = 0.05$ were performed between the responding and nonresponding patients for each time during treatment administration. Whenever there were multiple features computed due to, for example, changing existing parameters (e.g., different k and patch size values in texton method), the best feature selected most frequently, i.e., on most of the folds in LOSO classification by the SFS feature selection algorithm, was used for the statistical test of significance. Results are reported in Table III and arranged based on the dissimilarity measure.

Except for the texton-based approach using the MMD dissimilarity measure, the two classes for both MBF and AAC demonstrated almost no statistical significance. The proposed texton approach using MMD for the MBF, achieved statistically significant differences between the two classes for all the time groups in the study. Similarly, for the AAC, this same configuration achieved statistical significance only starting at week 4. The "MeanInt" features for both parameters demonstrated increasing statistical significance, more so for the MBF where the p values approached a significance threshold of 0.05. On the other hand, the ℓ_2 norm metric never identified a significant difference in changes between the two populations. The closest rival approach, i.e., the LBP features with MMD dissimilarities could only identify a significant difference in changes after week 4. An interesting observation is the consistent absence of statistical significance at week 1 for non-texton approaches, and as a whole, for the other two time groups (except for LBP-MMD), indicating that the texton approach is capable of differentiating between the two classes as early as week 1. In [36], an earlier study where their data set was a subset of our patient data set, in addition to textural biomarkers, it was demonstrated that using "MeanInt" for the MBF parameter achieved statistical significance starting at week 1, however, only five patients out of the 20 were nonresponders. As seen in the previous section, greater statistical significance correlated with a higher classification performance obtained using the proposed texton approach versus the other texture methods.

V. DISCUSSION AND CONCLUSION

The results of this study validated the viability of the developed CAT system based on using QUS parameters as biomarkers in conjunction with data-driven texture methods for volumetric data feature extraction. The different ultrasound characteristics of responding tumors compared to nonresponding tumors were noninvasively quantified as early as week 1, in patients with LABC receiving neoadjuvant chemotherapy when using the proposed CAT system based on the Texton-MMD approach.

Overall, the combination of a data-driven texture method and a kernel-based dissimilarly measure, i.e., Texton-MMD

method, resulted in the strongest performance amongst other feature-distance combinations, particularly on week 1 using the MBF spectral parametric map and on weeks 4 and 8 using the AAC BSC parameter. The improvement was specifically significant compared to the LBP-MMD approach on week 1, and the GLCM-MMD method on all weeks, which indicates the advantage of using a data-driven texture method over classical texture methods such as GLCM, and over texture methods relying on predefined operators such as LBPs. Moreover, the performance was significantly improved compared with the ℓ_2 norm, which demonstrated the importance of using kernel-based dissimilarity measures in calculating the distances between the "pretreatment" and "mid-treatment" scans, which was in line with our previous finding [14]. The significant improvement achieved by the proposed CAT system on weeks 1 and 4 validated its potential for assessing the effectiveness of treatment regimens in the early stages of neoadjuvant chemotherapy for LABC.

The earliest investigations on tumor classification using QUS biomarkers were able to successfully display the differences that arise in spectral parameters during cell death by using the mean of the parameters to represent an entire ROI [40], [41]. While the use of the mean of parametric maps does provide a means to interpret the status of the tumor cell over time, textural properties of parametric maps have shown a higher correlation to changes in the tumor tissue [10]. Specifically, texture properties derived from the GLCM, for example, contrast and homogeneity, have been demonstrated to have a higher performance than the "MeanInt" in response classification studies especially early on during a course of treatment [10], [36]. The reason for the improved classification performance when using more complex texture methods versus the simple mean method, may be due to the additional consideration of structures (manifested to pixel intensities) next to each other within a sufficiently sized neighborhood. This study confirmed the previous findings on a larger cohort of LABC patient data and using more advanced textural features.

The use of machine learning algorithms has also been accounted for in computer-aided diagnosis, showing high performance in CT [17], [42], [43] and MRI images [44]. We provided a novel study on tumor response classification of QUS parametric maps using a texton learning approach, presenting a comparison of feature extraction from intrinsic textural features versus predefined texture patterns. Predefined patterns including LBP have shown promising performance in image analysis; however, in our study, they did not perform as well as the raw-pixel intensity texton approach. One main reason could be due to the fact that the texton-based method is a dictionary-based approach, constructing one dictionary on all the scans of a patient, which accounts for relative information of scans (and hence, volumetric information of tumors) in the subsequent feature models extraction from individual scans. The LBPs, in contrast, extract feature models from each scan using predefined generic binary operators (not learned from the data), hence, not accounting for the volumetric information in model construction. Another explanation could be due to the fact that the LBPs are, by design, invariant to gray-scale intensities. While this can be an attractive attribute in some applications, where the designed system should be invariant to changes in illumination, in cancer response monitoring using QUS methods, gray-scale intensity conveys important information. For example, in MBF parametric map, it represents the ultrasound backscatter power, which increases with different modalities of cell death [7]. Therefore, excluding this information can deteriorate the performance of a CAT system. Texton-based methods based on raw-pixel representation, however, can account for this information and intrinsically lead to better classification results compared to the LBPs. While this can be fixed for the LBPs by, e.g., combining them with the histograms of intensity as shown in [14], it adds to the complexity of the approach, whereas textons do not need this additional step in feature design.

The novel engineered texton method was used here for the application of cancer response monitoring to build one dictionary of textons for each patient at each time interval during treatment. There are several major advantages for building the dictionary using the proposed method: first, since in a CAT system comparison between the scans from "pretreatment" and "mid-treatment" is one major component, building the dictionary in the aforementioned manner makes it more representative of the two scans types, and subsequently leads to better classification results. Second, the approach learns the dictionary irrespective of whether the patients are in the training or tests sets. Therefore, as new patients are coming, the dictionary can easily be updated by building a dictionary merely on the scans from this patient. This significantly saves the computation time for building the dictionary as there is no need to update the dictionary for all patients when the train and test sets are changed, e.g., due to cross-validation or due to adding new patients.

The Texton-based method is a dictionary learning and sparse representation [45] method specifically tailored to texture images. However, in the texton-based method, each patch in a texture image is represented by only one single texton in the dictionary (the closest match). This is a kind of sparse representation, in which only one atom in the dictionary is active per patch. However, this might not be an ideal model and using other dictionary learning and sparse representation methods that can include more than one dictionary atom in modeling each patch in texture image can potentially improve the classification performance [46]. Especially, supervised dictionary learning methods [47], [48] that learn one subdictionary per "pretreatment" and another subdictionary per "mid-treatment" scans are of interest in this application, and research in this direction is currently underway in our group.

In this paper, one parametric map was applied to the proposed CAT system at a time. However, it is certainly possible to use complementary information in the parametric maps by combining them in the proposed pipeline. Combining the parametric maps can be performed at several different stages: 1) at feature level; 2) after the computation of the dissimilarity metrics; and 3) at decision level by employing multiple classifier systems (MCSs) [49]. At feature level, in addition, combining can be accomplished in many different ways, e.g., by simply feature fusion or by using advanced

multiview learning [50] approaches such as multiview supervised dictionary learning [51]. At the decision level, there are many structures for MCSs as well as different combining rules such as majority voting, mean, product, min, or max combining rules. Moreover, there will be a question on how many different parametric maps should be combined, e.g., whether we just combine the AAC and MBF or include more spectral and BSC parametric maps. Combining parametric maps will be addressed in our future research.

In conclusion, the proposed CAT system provides promise that QUS methods in conjunction with data-driven texture analysis and advanced machine learning techniques to spur advances in cancer treatment in the near future by providing a means for the early detection of cell death. These methods play an integral role in association with other imaging modalities such as MRI and PET as treatment becomes more personalized and moves to the molecular level.

REFERENCES

[1] L. A. Newman, "Epidemiology of locally advanced breast cancer," *Seminars Radiat. Oncol.*, vol. 19, no. 4, pp. 195–203, 2009.

[2] S. H. Giordano, "Update on locally advanced breast cancer," *Oncologist*, vol. 8, no. 6, pp. 521–530, 2003.

[3] M. De Lena, R. Zucali, G. Viganotti, P. Valagussa, and G. Bonadonna, "Combined chemotherapy-radiotherapy approach in locally advanced (T_3b-T_4) breast cancer," *Cancer Chemother. Pharmacol.*, vol. 1, no. 1, pp. 53–59, 1978.

[4] M. E. Juweid and B. D. Cheson, "Positron-emission tomography and assessment of cancer therapy," *New Engl. J. Med.*, vol. 354, no. 5, pp. 496–507, 2006.

[5] T. H. Witney and K. M. Brindle, "Imaging tumour cell metabolism using hyperpolarized ^{13}C magnetic resonance spectroscopy," *Biochem. Soc. Trans.*, vol. 38, no. 5, pp. 1220–1224, 2010.

[6] O. Falou *et al.*, "Diffuse optical spectroscopy evaluation of treatment response in women with locally advanced breast cancer receiving neoadjuvant chemotherapy," *Transl. Oncol.*, vol. 5, no. 4, pp. 238–246, 2012.

[7] G. J. Czarnota and M. C. Kolios, "Ultrasound detection of cell death," *Imag. Med.*, vol. 2, no. 1, pp. 17–28, 2010.

[8] M. J. Gangeh, A. Sadeghi-Naini, M. Diu, M. S. Kamel, and G. Czarnota, "Categorizing extent of tumour cell death response to cancer therapy using quantitative ultrasound spectroscopy and maximum mean discrepancy," *IEEE Trans. Med. Imag.*, vol. 33, no. 6, pp. 1390–1400, Jun. 2014.

[9] E. J. Feleppa, J. Mamou, C. R. Porter, and J. Machi, "Quantitative ultrasound in cancer imaging," *Seminar Oncol.*, vol. 38, no. 1, pp. 136–150, 2011.

[10] A. Sadeghi-Naini *et al.*, "Conventional frequency ultrasonic biomarkers of cancer treatment response *in vivo*," *Transl. Oncol.*, vol. 6, no. 3, pp. 234–243, Jun. 2013.

[11] A. Sadeghi-Naini *et al.*, "Quantitative evaluation of cell death response *in vitro* and *in vivo* using conventional-frequency ultrasound," *Oncoscience*, vol. 2, no. 8, pp. 716–726, 2015.

[12] M. L. Oelze and J. Mamou, "Review of quantitative ultrasound: Envelope statistics and backscatter coefficient imaging and contributions to diagnostic ultrasound," *IEEE Trans. Ultrason. Ferroelect. Freq. Control.*, vol. 63, no. 2, pp. 336–351, Feb. 2016.

[13] G. J. Czarnota *et al.*, "Ultrasound imaging of apoptosis: High-resolution non-invasive monitoring of programmed cell death *in vitro*, *in situ* and *in vivo*," *Brit. J. Cancer*, vol. 81, no. 3, pp. 520–527, 1999.

[14] M. J. Gangeh, H. Tadayyon, L. Sannachi, A. Sadeghi-Naini, W. T. Tran, and G. J. Czarnota, "Computer aided theragnosis using quantitative ultrasound spectroscopy and maximum mean discrepancy in locally advanced breast cancer," *IEEE Trans. Med. Imag.*, vol. 35, no. 3, pp. 778–790, Mar. 2016.

[15] S. D. Rice *et al.*, "Analysis of chemotherapeutic response heterogeneity and drug clustering based on mechanism of action using an *in vitro* assay," *Anticancer Res.*, vol. 30, no. 7, pp. 2805–2811, 2010.

[16] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2032–2047, Nov. 2009.

[17] M. J. Gangeh, L. Sørensen, S. B. Shaker, M. S. Kamel, M. de Bruijne, and M. Loog, "A texton-based approach for the classification of lung parenchyma in CT images," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Berlin, Germany: Springer-Verlag, 2010, pp. 595–602.

[18] L. Sørensen, M. J. Gangeh, S. B. Shaker, and M. de Bruijne, "Texture classification in pulmonary CT," in *Lung Imaging and Computer Aided Diagnosis*, A. El-Baz and J. S. Sure, Eds. Boca Raton, FL, USA: CRC Press, 2011, pp. 343–367.

[19] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[20] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[21] M. J. Gangeh *et al.*, "Cancer response monitoring using texton-based approach in locally advanced breast cancer," in *Proc. 40th Int. Symp. Ultrason. Imag. Tissue Characterization*, 2015.

[22] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vis.*, vol. 62, nos. 1–2, pp. 61–81, 2005.

[23] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[24] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, no. 5802, pp. 91–97, 1981.

[25] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 29–44, Feb. 2001.

[26] O. G. Cula and K. J. Dana, "3D texture recognition using bidirectional feature histograms," *Int. J. Comput. Vis.*, vol. 59, pp. 33–60, Aug. 2004.

[27] D. Mease, A. J. Wyner, and A. Buja, "Boosted classification trees and class probability/quantile estimation," *J. Mach. Learn. Res.*, vol. 8, pp. 409–439, May 2007.

[28] R. C. Holte, L. E. Acker, and B. W. Porter, "Concept learning and the problem of small disjuncts," in *Proc. 11th Int. Joint Conf. Artif. Intell.* San Mateo, CA, USA: Morgan Kaufmann, 1989, pp. 813–818.

[29] G. M. Clarke, S. Eidt, L. Sun, G. Mawdsley, J. T. Zubovits, and M. J. Yaffe, "Whole-specimen histopathology: A method to produce whole-mount breast serial sections for 3-D digital histopathology imaging," *Histopathology*, vol. 50, no. 2, pp. 232–242, 2007.

[30] E. Eisenhauer *et al.*, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *Eur. J. Cancer*, vol. 45, no. 2, pp. 228–247, 2009.

[31] H. Tadayyon, A. Sadeghi-Naini, L. Wirtzfeld, F. C. Wright, and G. Czarnota, "Quantitative ultrasound characterization of locally advanced breast cancer by estimation of its scatterer properties," *Med. Phys.*, vol. 41, no. 1, p. 012903, 2014.

[32] F. L. Lizzi, M. Astor, T. Liu, C. Deng, D. J. Coleman, and R. H. Silverman, "Ultrasonic spectrum analysis for tissue assays and therapy evaluation," *Int. J. Imag. Syst. Technol.*, vol. 8, no. 1, pp. 3–10, 1997.

[33] L. X. Yao, J. A. Zagzebski, and E. L. Madsen, "Backscatter coefficient measurements using a reference phantom to extract depth-dependent instrumentation factors," *Ultrason. Imag.*, vol. 12, no. 1, pp. 58–70, 1990.

[34] M. F. Insana and T. J. Hall, "Parametric ultrasound imaging from backscatter coefficient measurements: Image formation and interpretation," *Ultrason. Imag.*, vol. 12, no. 4, pp. 245–267, 1990.

[35] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. TSMC-3, no. 6, pp. 610–621, Nov. 1973.

[36] A. Sadeghi-Naini *et al.*, "Early prediction of therapy responses and outcomes in breast cancer patients using quantitative ultrasound spectral texture," *Oncotarget*, vol. 5, no. 11, pp. 3497–3511, 2014.

[37] L. Sannachi *et al.*, "Non-invasive evaluation of breast cancer response to chemotherapy using quantitative ultrasonic backscatter parameters," *Med. Image. Anal.*, vol. 20, no. 1, pp. 224–236, 2015.

[38] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.

[39] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1601–1608.

[40] B. Banihashemi, R. Vlad, B. Debeljevic, A. Giles, M. C. Kolios, and G. J. Czarnota, "Ultrasound imaging of apoptosis in tumor response: Novel preclinical monitoring of photodynamic therapy effects," *Cancer Res.*, vol. 68, no. 20, pp. 8590–8596, 2008.
[41] R. M. Vlad, S. Brand, A. Giles, M. C. Kolios, and G. J. Czarnota, "Quantitative ultrasound characterization of responses to radiotherapy in cancer mouse models," *Clin. Cancer Res.*, vol. 15, no. 6, pp. 2067–2075, 2009.
[42] M. J. Gangeh, L. Sørensen, S. B. Shaker, M. S. Kamel, and M. de Bruijne, "Multiple classifier systems in texton-based approach for the classification of CT images of lung," in *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*. Berlin, Germany: Springer-Verlag, 2010, pp. 153–163.
[43] E.-L. Chen, P.-C. Chung, C.-L. Chen, H.-M. Tsai, and C.-I. Chang, "An automatic diagnostic system for CT liver image classification," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 6, pp. 783–794, Jun. 1998.
[44] E. I. Zacharaki *et al.*, "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme," *Magn. Reson. Med.*, vol. 62, no. 6, pp. 1609–1618, Dec. 2009.
[45] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[46] M. J. Gangeh, A. Ghodsi, and M. S. Kamel, "Dictionary learning in texture classification," in *Proc. 8th Int. Conf. Image Anal. Recognit.* Berlin, Germany: Springer-Verlag, 2011, pp. 335–343.
[47] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1033–1040.
[48] M. J. Gangeh, A. Ghodsi, and M. S. Kamel, "Kernelized supervised dictionary learning," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4753–4767, Oct. 2013.
[49] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ, USA: Wiley, 2004.
[50] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2031–2038, 2013.
[51] M. J. Gangeh, P. Fewzee, A. Ghodsi, M. S. Kamel, and F. Karray, "Multiview supervised dictionary learning in speech emotion recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 6, pp. 1056–1068, Jun. 2014.

Mehrdad J. Gangeh (M'05–SM'15) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2013.

He was a Research Scientist at the Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada, and the Department of Radiation Oncology, Sunnybrook Health Sciences Center, Toronto. He is currently a Senior Artificial Intelligence Scientist with the Global Innovation Team, EY, Palo Alto, CA, USA. He is also an Adjunct Professor with the Department of System Design Engineering, University of Waterloo. His research interests are multiview learning, dictionary learning and sparse representation, kernel methods, and deep learning with applications to natural language processing, text analytics, medical imaging, pattern recognition, big data analytics, computer vision, data mining, and bioinformatics.

Dr. Gangeh is an Associate Editor of the IEEE TRANSACTIONS ON MEDICAL IMAGING.

Simon Liu received the bachelor's degree in chemical engineering with an impending entrance into the Master's program for materials engineering.

During numerous internships, he has been able to work in various fields, including cancer research, lithium and sodium ion batteries, liquid metal batteries, and reflective displays. His future aspirations lie in chemistry and materials synthesis for sustainable and environmentally responsible technologies. This mainly encompasses alternative energy and green building materials. In the far future, he aims to become a teacher due to his passion for mentoring and guiding young minds.

Hadi Tadayyon received the bachelor's degree in electrical and computer engineering from Western University, London, ON, Canada, in 2008, the master's degree in electrical and computer engineering from Queen's University, Kingston, ON, Canada, in 2010, and the Ph.D. degree in medical biophysics from the University of Toronto, Toronto, ON, Canada, in 2015.

His research interests include ultrasound tissue characterization, medical image analysis in conjunction with machine learning, and cancer imaging biomarkers.

Gregory J. Czarnota is currently a Senior Scientist with the Imaging Division and an M.D. with the Department of Radiation Oncology, Sunnybrook Health Sciences Centre, Toronto, ON, Canada, with applied research in breast cancer patients. His laboratory is conducting research focused on using ultrasound imaging and spectroscopy at conventional and high frequencies to detect apoptosis and other forms of cell death in response to chemotherapy and radiation therapy. His basic-science research interests include studies in biochemistry, chromatin biology, biophysics, medicine, and oncology. His research is pertinent to the application here as along with Dr. M. Kolios, he discovered the use of QUS to detect cell death, demonstrated this in cell culture and animal models, and has been translating the results of monitoring human patients receiving chemotherapy with this technology.