Breast Cancer Detection With Low-Dimensional Ordered Orthogonal Projection in Terahertz Imaging

Tanny Chavez, *Student Member, IEEE*, Nagma Vohra, *Student Member, IEEE*, Jingxian Wu, *Senior Member, IEEE*, Keith Bailey, and Magda El-Shenawee, *Senior Member, IEEE*

***Abstract*—This article proposes a new dimension reduction algorithm based on low-dimensional ordered orthogonal projection, which is used for cancer detection with terahertz (THz) images of freshly excised human breast cancer tissues. A THz image can be represented by a data cube with each pixel containing a high-dimensional spectrum vector covering several THz frequencies, where each frequency represents a different dimension in the vector. The proposed algorithm projects the high-dimensional spectrum vector of each pixel within the THz image into a low-dimensional subspace that contains the majority of the unique features embedded in the image. The low-dimensional subspace is constructed by sequentially identifying its orthonormal basis vectors, such that each newly chosen basis vector represents the most unique information not contained by existing basis vectors. A multivariate Gaussian mixture model is used to represent the statistical distributions of the low-dimensional feature vectors obtained from the proposed dimension reduction algorithm. The model parameters are iteratively learned by using unsupervised learning methods, such as Markov chain Monte Carlo or expectation maximization, and the results are used to classify the various regions within a tumor sample. Experiment results demonstrate that the proposed method achieves apparent performance improvement in human breast cancer tissue over existing approaches such as one-dimensional Markov chain Monte Carlo. The results confirm that the dimension reduction algorithm presented in this article is a promising technique for breast cancer detection with THz images, and the classification results present a good correlation with respect to the histopathology results of the analyzed samples.**

***Index Terms*—Breast cancer, expectation maximization (EM), Gaussian mixture model (GMM), Gibbs sampling, low-dimensional ordered orthogonal projection (LOOP), terahertz (THz).**

I. Introduction

BREAST CANCER is one of the most common types of cancer among women with over two million new cases in 2018 [1]. Breast conserving surgery, also known as lumpectomy, is a commonly suggested treatment option when breast cancer is detected at an early stage. The aim of lumpectomy is to excise all the cancerous tissues surrounded by a small margin of healthy breast tissue [2]. Currently, the success of lumpectomy is determined through histopathology analysis of the excised tissue, which may take around ten days to process. As a result, one in five patients has to go under a second surgery to extract remaining cancerous tissues [3]. This necessitates the design of new technologies that can examine the margins of the freshly excised breast cancer tissue in the operation room while the surgery is still ongoing. In this context, terahertz (THz) imaging has shown promising results for tissue classification within freshly excised breast cancer tumors [4]–[8].

THz imaging has been used for various medical applications, such as the evaluation of brain injuries [9], colon cancer inspection [10], diagnosis of oral lichen planus [11], liver cancer identification [12], [13], breast cancer detection [4]–[8], etc. Different approaches are adopted by these works to identify the regions of interests from the rest of the sample, and the classifications are achieved by utilizing the distinguishing features of different regions embedded in THz signals. For instance, the electromagnetic propagation parameters, such as absorption coefficient, complex permittivity, refractive index, and dielectric loss tangent of the cells, are used as features for the detector of colon cancer [10]. Many studies employ statistical learning and machine learning techniques to achieve THz image segmentation. An unsupervised k-means clustering method with ranked set sampling is proposed in [14] for the segmentation of THz images. Supervised learning techniques in THz imaging include support vector machines (SVMs) [11]–[13], [15], probabilistic neural networks [12], [13], and deep neural networks [16]. While machine learning techniques have proven to achieve good correlation with respect to their pathology counterparts, the need for a large amount of training samples make their applications complicated and occasionally inconsistent.

A THz image can be represented by a data cube with each pixel containing a high-dimensional spectrum vector covering several THz frequencies, where each frequency represents a different dimension in the vector. The high-dimensional vector per pixel contains both common features that are shared by all regions within a tissue sample, and unique features that can be used to distinguish different regions. Thus, it is desirable to extract the unique features embedded in the THz signals to reduce complexity and improve accuracy. In [5] and [8], the high-dimensional THz waveform per pixel is summarized into

Manuscript received July 30, 2019; revised October 25, 2019; accepted December 6, 2019. Date of publication December 24, 2019; date of current version March 3, 2020. This work was supported in part by the National Institutes of Health under Award R15CA208798 and in part by the National Science Foundation under Award 1408007 and Award 1711087 and 1711087. *(Corresponding author: Tanny Chavez.)*

T. Chavez, N. Vohra, J. Wu, and M. El-Shenawee are with the Department of Electrical Engineering, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: tachavez@email.uark.edu; nvohra@email.uark.edu; wuj@uark.edu; magda@uark.edu).

K. Bailey is with the Veterinary Diagnostic Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61802 USA (e-mail: kbailey1@illinois.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TTHZ.2019.2962116

a scalar, such as the peak of the reflected time-domain signal or the energy over a certain frequency band. The one-dimensional (1-D) feature extractions used in [5] and [8] show good performance for tumor samples with two regions, but its performance drops considerably when there are three or more regions in the sample tissue. While some studies summarize the information per pixel using a pre-established characteristic [10], the usage of dimension reduction algorithms has gained interest due to their systematic information extraction capabilities. Some commonly used dimension reduction algorithms are principal component analysis (PCA) [11], [12], [15], Isomap [12], [13], and linear preserving projections [12].

In this article, we propose a new low-dimensional ordered orthogonal projection (LOOP) algorithm, which is used to extract low-dimensional features embedded in the high-dimensional THz waveform. The low-dimensional feature extraction is achieved by projecting the original THz signal into a low-dimensional subspace containing the majority of the salient information necessary for classification. The low-dimensional subspace is constructed by sequentially identifying its orthonormal basis vectors with a specific order, such that each new basis vector is chosen as the one that contains the most amount of unique information not represented by all previous basis vectors. Such an approach can ensure that all spectrum vectors within the dataset can be represented by the basis vectors with minimum information loss; thus, the majority of the useful information in the original THz signal is captured by the constructed subspace. Unlike single-dimensional feature extraction methods that are limited by the selection of one physical parameter of the THz signal [5], [8], the LOOP algorithm extracts the most significant information from the waveform as a low-dimensional vector, which represents a combination of all important features. The elements in the low-dimensional vector do not correspond to a specific physical feature, and they are usually combinations of several important physical features. While an early version of this dimension reduction algorithm was briefly discussed in [17], the LOOP algorithm presented in this article explores a new ordering technique that differs significantly from the projection method in [17]. In addition, the work presented in [17] was focused on murine samples, while the results presented in this article focus on human breast tumor samples.

The low-dimensional feature vector is analyzed and modeled by using a multivariate Gaussian mixture model (GMM) [18], with each component in the GMM corresponding to one possible tissue type within the sample. The prevalence of different tissue types within a sample is modeled by using the weight or prior probability for each component in the GMM. Such a probabilistic approach can capture the statistical nature of the THz signal and provide important reliability information that is not available in deterministic approaches. Two unsupervised learning algorithms, Markov chain Monte Carlo (MCMC) [19] and expectation maximization (EM) [20], are used to learn the parameters of the GMM with the low-dimensional feature vectors. Given that the acquisition of breast cancer samples is limited and laborious, in particular for fresh human samples, unsupervised learning algorithms are preferred due to the lack of a training phase. The results are used to classify different regions within sample tissues. Unlike existing works that focus on the binary classification of a tissue (cancerous versus healthy tissue) [12], this article focuses on the identification of different regions, such as collagen, fibro, and fat, within heterogeneous breast cancer samples. The proposed LOOP algorithm with unsupervised learning is applied to THz imaging of freshly excised human breast cancer tissue with three regions: cancer, collagen or fibro, and fat. Experiment results demonstrated that the proposed LOOP algorithm is a promising technique for cancer detection with THz images, and the classification results present a good correlation with respect to results obtained from histopathology analysis.

The rest of this article is organized as follows. Section II presents the experiment setup and data collection process. Section III introduces the problem formulation and notations used in this article. Details of the LOOP algorithm are explained in Section IV. Section V defines the GMM and the unsupervised learning algorithms based on the low-dimensional vector obtained by LOOP. Section VI shows the experimental results. Section VII concludes this article.

II. Experiment Setup

The tissue samples handled in this article follow the Environmental Health and Safety Protocol of the University of Arkansas. The experimental work done in this article makes use of human breast cancer tissues #ND10898, ND15526, and ND15588 obtained from the National Disease Research Interchange within 24 h of excision. These samples were obtained via left breast masectomy from a 59-year-old patient diagnosed with III/III grade infiltrating dual carcinoma (IDC), radical masectomy from a 90-year-old patient with III/III grade IDC, and masectomy from a 63-year-old patient with II/III grade IDC, respectively. On receiving the tissue in the THz laboratory, it was removed from the Dulbecco's Modified Eagle Medium (DMEM) solution [see Fig. 1(a)]. After removing excess water using filter paper [see Fig. 1(b)], the tissue was positioned between two polystyrene plates and pressed softly to make the imaging surface as flat as possible, while also maintaining the original shape of the tissue [see Fig. 1(c)]. This arrangement of the tissue was then mounted on the scanner stage for the reflection imaging procedure, as shown in Fig. 1(d).

The reflection measurements were taken by using a TPS Spectra 3000 pulsed THz imaging and spectroscopy system (from TeraView Ltd., U.K.). The diagram of the system is shown in Fig. 2(a). The system uses a Ti:Sapphire laser that produces an 800-nm pulse to excite the THz emitter and the THz receiver. Upon excitation, the THz emitter generates a time-domain THz pulse, as shown in Fig. 2(b). The Fourier transform of the pulse, as shown in Fig. 2(c), demonstrates a power spectrum of pulse ranging from 0.1 to 4 THz. This emitted pulse is made incident on the sample through a set of mirrors, and the reflected pulse from the sample is directed toward the THz receiver [8]. In the reflection mode measurements, both the THz emitter and the detector are offset 30° with respect to the normal direction on the sample. To obtain the THz-reflected signal at each pixel on the tissue to produce an image, the scanning stage was set to move in increments of 200-μm step size using a stepper motor.

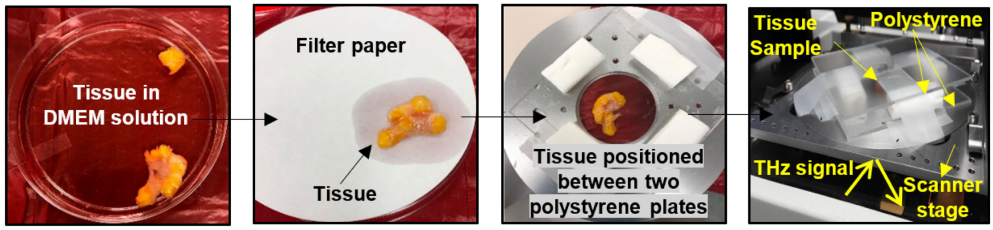

Fig. 1. Sample preparation process. (a) Tissue immersed in DMEM solution. (b) Removal of excess water in tissue using filter paper. (c) Tissue positioned in a sandwich between two polystyrene plates. (d) Positioning the tissue sandwich on the scanner stage for imaging.

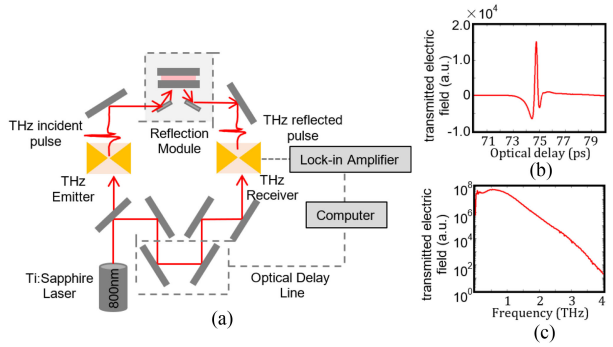

Fig. 2. (a) THz system diagram for reflection imaging. (b) Incident time-domain THz pulse. (c) Frequency spectrum of THz pulse in (b).

The total time span of the imaging process was ∼30–40 min. During this time, the samples could get slightly dried on the surface; however, the pathologist did not report any damage at the cellular level. For imaging, we focus the THz beam on the tissue surface and conduct two scans; the first one is a quick line scan using 400 μm to assure the flat level of the tissue based on the B-scan (cross section), and the second scan is for the final image in the xy plane taken at 200-μm step size. Upon finishing the scanning process, the tissue was immersed in formalin solution and shipped to the Oklahoma Animal Disease Diagnostic Laboratory for the pathology process. The histopathology process involves fixing the tissue in formalin and embedding it in paraffin blocks. Furthermore, from the formalin fixed paraffin embedded (FFPE) tissue blocks, two ∼3–4-μm-thick slices were cut, stained with hematoxylin and eosin, and fixed on the glass slides to produce pathology images using low-power microscope. For assessing the images of the freshly excised tumor and the FFPE tissue block, the THz images are compared with the pathology images, as will be discussed in Section VI.

III. Problem Formulation

The problem formulation and notations are described in this section. Let the tensor $\boldsymbol{\mathcal{W}} \in \mathcal{R}^{N_1 \times N_2 \times T}$ represent the THz image of size $N_1 \times N_2$. Each pixel $\boldsymbol{\mathcal{W}}_{n_1,n_2,*}$ corresponds to the reflected time-domain signal, which contains T time samples at the output of the THz system. The subscripts $n_1 \in \{1, \ldots, N_1\}$ and $n_2 \in \{1, \ldots, N_2\}$ represent the coordinates of the pixel along the x and y axes, respectively.

For simplicity, the tensor $\boldsymbol{\mathcal{W}}$ is unfolded into a matrix $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N] \in \mathcal{R}^{T \times N}$ with $N = N_1 \times N_2$, such that each column of $\boldsymbol{W}$ represents the T time samples of one pixel of the THz image. Once unfolded, the algorithm computes the complex spectrum of the signal per pixel in the frequency domain by using fast Fourier transform (FFT). The frequency-domain representation of the ith pixel is $\boldsymbol{y}_i = \mathcal{F}(\boldsymbol{w}_i)$, where $\mathcal{F}(\cdot)$ is the FFT operator. Since $\boldsymbol{w}_i$ is real, the FFT of $\boldsymbol{w}_i$ is even-symmetric. Thus, the size of $\boldsymbol{y}_i$ is $F = \frac{T}{2}$. Define the frequency-domain THz image matrix as $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N] \in \mathcal{C}^{F \times N}$.

In our experiment setup, each pixel contains $N = 1024$ time samples with a sampling period $T_0 = 0.026$ ps. Correspondingly, the frequency-domain representation of each pixel has 512 frequency samples. Theoretically, the frequency span of each pixel is $\frac{1}{2T_0} = 18.97$ THz, with the frequency-domain resolution being $F_0 = \frac{1}{NT_0} = 37.05$ GHz. Considering the physical limitations of the THz system, the frequency-domain signal of each pixel is limited to [0.1, 4] THz, which corresponds to the system's operation range. Therefore, the number of frequency samples per pixel is reduced to $F = 106$.

Either the original complex THz spectrum or its amplitude can be used to classify the various regions inside a tissue sample. The subsequent analysis is applicable to both the complex spectrum and the amplitude spectrum. To unify notations, define a new spectrum matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]$ to represent both the complex and amplitude spectra. For analysis of the complex spectrum, we have $\boldsymbol{Y} = \boldsymbol{D}$; for analysis of the amplitude spectrum, $\boldsymbol{Y}$ is obtained by replacing all elements in $\boldsymbol{D}$ with their respective amplitudes.

We will perform cancer detection by utilizing the frequency-domain THz matrix $\boldsymbol{Y}$, such that each pixel can be classified into one category from a finite set of tissue types, such as cancer, fat, muscle, etc. The information of the ith pixel is represented by the frequency-domain vector $\boldsymbol{y}_i$, which has a relatively large dimension of $F = 106$. The frequency-domain vector $\boldsymbol{y}_i$ contains both common features that are shared by multiple tissue types and unique features that can be used to distinguish different types of tissues. Performing classification directly over $\boldsymbol{y}_i$ means that the algorithm needs to process both common features and unique features. This will incur unnecessarily high computation complexity, and the overall performance of the classifier will be negatively affected by the Hughes phenomenon [21].

It is thus desirable to perform low-dimensional feature extraction before classification. With low-dimensional feature extraction, the high-dimensional vector $\boldsymbol{y}_i$ can be mapped to a low-dimensional feature domain that contains the majority of the salient information of the unique features. Such an approach can significantly improve the classification accuracy and efficiency.

IV. Low-Dimensional Ordered Orthogonal Projection

In this section, we propose a LOOP algorithm to achieve low-dimensional feature extraction from the frequency-domain THz matrix $\boldsymbol{Y}$.

The main objective of the algorithm is to identify a low-dimensional subspace of the space spanned by the columns of $\boldsymbol{Y}$, and the subspace should contain the majority of the salient information of the unique features embedded in $\boldsymbol{Y}$. Once the subspace is identified, the frequency-domain vector of each pixel can then be projected into the subspace to achieve low-dimensional feature extraction.

The subspace can be described by an orthonormal basis $\mathcal{B}_L = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L\}$, where $L < F$ is the dimension of the subspace. The LOOP algorithm identifies $\mathcal{B}$ by using a modified Gram–Schmidt (GS) process [22]. The conventional GS process sequentially identifies a set of orthonormal vectors that form the basis of the space spanned by a set of vectors. The sequential procedure of the conventional GS is performed in an arbitrary order without considering the features embedded in the vectors. The LOOP algorithm improves the GS process by ordering the sequentially identified orthonormal basis vectors, such that most of the unique features embedded in $\boldsymbol{Y}$ are contained in the subspace spanned by the first L orthonormal basis vectors.

To achieve this goal, the LOOP algorithm calculates each new orthonormal basis vector by using the pixel that is least represented by all previous basis vectors. That is, each new orthonormal basis vector is chosen as the one that contains the most amount of unique information not represented by all previous basis vectors. Following such an ordered sequential process, most of the unique information embedded in $\boldsymbol{Y}$ is captured by the first few basis vectors. Details of the LOOP algorithm are described as follows.

In the LOOP algorithm, the first orthonormal basis vector is calculated by normalizing the average vector of all pixels as

$$\boldsymbol{b}_1 = \frac{\bar{\boldsymbol{y}}}{\|\bar{\boldsymbol{y}}\|} \tag{1}$$

where $\bar{\boldsymbol{y}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i$, and $\|\bar{\boldsymbol{y}}\| = \sqrt{\bar{\boldsymbol{y}}^H \bar{\boldsymbol{y}}}$ is the norm of $\bar{\boldsymbol{y}}$, with $\bar{\boldsymbol{y}}^H$ being the vector conjugate transpose operator.

The subsequent orthonormal basis vectors are calculated in a sequential manner. Assume that the first l orthonormal basis vectors have been identified, and they are represented as $\mathcal{B}_l = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_l]$. The $(l+1)$th basis vector will be calculated by using the pixel that is least represented by $\boldsymbol{B}_l$. How well a vector is represented in a subspace can be measured by using the angle between the vector and its projection in the subspace. A right angle means the subspace does not contain any information of the vector, and a 0° angle means the vector can be fully represented by the subspace.

The projection of the vector $\boldsymbol{y}_i$ onto a subspace spanned by $\mathcal{B}_l$ can be calculated as

$$P_{\mathcal{B}_l}(\boldsymbol{y}_i) = \sum_{j=1}^{l} \langle \boldsymbol{y}_i, \boldsymbol{b}_j \rangle \boldsymbol{b}_j \tag{2}$$

and $\langle \boldsymbol{y}_i, \boldsymbol{b}_j \rangle = \boldsymbol{y}_i^H \boldsymbol{b}_j$ is the inner product between vectors $\boldsymbol{y}_i$ and $\boldsymbol{b}_j$.

Denote the angle between the two vectors $\boldsymbol{y}_i$ and $P_{\mathcal{B}_l}(\boldsymbol{y}_i)$ as $\theta_{i,l} = \angle(\boldsymbol{y}_i, P_{\boldsymbol{B}_l}(\boldsymbol{y}_i))$; then, we have

$$\cos(\theta_{i,l}) = \frac{\langle \boldsymbol{y}_i, P_{\mathcal{B}_l}(\boldsymbol{y}_i) \rangle}{\|\boldsymbol{y}_i\| \cdot \|P_{\mathcal{B}_l}(\boldsymbol{y}_i)\|}. \tag{3}$$

Based on the above notations, we can identify the pixel that is least represented by the subspace $\mathcal{B}_l$ as

$$\boldsymbol{u}_{l+1} = \underset{\boldsymbol{y}_i \in \mathcal{Y}_l}{\operatorname{argmin}} \, |\cos(\theta_{i,l})| \tag{4}$$

where $\mathcal{Y}_l$ contains all the $\boldsymbol{y}_i$ vectors that are not in the subspace spanned by $\mathcal{B}_l$.

Once the vector $\boldsymbol{u}_{l+1}$ is identified, the $(l+1)$th basis vector, $\boldsymbol{b}_{l+1}$, can then be calculated by following the GS procedure:

$$\boldsymbol{v}_{l+1} = \boldsymbol{u}_{l+1} - P_{\mathcal{B}_l}(\boldsymbol{u}_{l+1}) \tag{5}$$

$$\boldsymbol{b}_{l+1} = \frac{\boldsymbol{v}_{l+1}}{\|\boldsymbol{v}_{l+1}\|}. \tag{6}$$

The procedure is repeated until $|\min_i \cos(\theta_{i,l})|$ is less than a predefined threshold or a predefined dimension L is reached. Once the orthonormal basis $\mathcal{B}_L$ is identified, we can project each pixel into the subspace spanned by $\mathcal{B}_L$ to achieve a low-dimensional representation of the THz image. Define $\boldsymbol{B}_L = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L] \in \mathcal{C}^{F \times L}$; then, the low-dimensional representation of $\boldsymbol{y}_i$ can be expressed as

$$\boldsymbol{y}_i = \boldsymbol{B}_L \times \boldsymbol{z}_i, \text{ for } i = 1, \ldots, N. \tag{7}$$

The output of the LOOP algorithm is the low-dimensional representation of the THz image in the feature subspace as $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N] \in \mathcal{C}^{L \times N}$, and it can also be represented in a compact form as

$$\boldsymbol{Y} = \boldsymbol{B}_L \times \boldsymbol{Z} \tag{8}$$

where $\boldsymbol{Z}$ can be determined using a least-squares approach.

V. Unsupervised Learning With the GMM

In this section, we present two unsupervised learning methods to classify the pixels based on the low-dimensional feature matrix $\boldsymbol{Z}$. Both methods are developed by using GMMs.

In the complex spectrum analysis, the elements in $\boldsymbol{Z}$ are complex numbers. To simplify notation, define a real-valued matrix by separating the real and imaginary part of $\boldsymbol{Z}$ as [23]

$$\boldsymbol{X} = [\Re(\boldsymbol{Z}^T), \Im(\boldsymbol{Z}^T)]^T \in \mathcal{R}^{2L \times N}. \tag{9}$$

On the other hand, for the amplitude spectrum analysis, all elements in $\boldsymbol{Z}$ are real numbers, and we define $\boldsymbol{X} = \boldsymbol{Z} \in \mathcal{R}^{L \times N}$.

The ith column of $\boldsymbol{X}$ is denoted by $\boldsymbol{x}_i$. In the GMM, it is assumed that the low-dimensional feature vector $\boldsymbol{x}_i$ follows a multimodal Gaussian distribution, with each mode corresponding to a specific region within the sample tissue. The GMM can be represented as

$$f\left(\boldsymbol{x}_i | [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, q_k]_{k=1}^{K}\right) = \sum_{k=1}^{K} q_k g(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{10}$$

where K is the number of categories in the sample tissue, q_k is the prior probability of a pixel in the kth category, and

$g(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian probability density function with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

Define a set of latent variables, $\zeta_i \in \{1, \ldots, K\}$, which are used to indicate the classification result of the ith pixel, for $i = 1, \ldots, N$. That is, $\zeta_i = k$ indicates that the ith pixel belongs to the kth category. It is assumed that the latent variable ζ_i follows a multinomial distribution with prior probability $\pi(\zeta_i = k) = q_k$, for $k = 1, \ldots, K$.

The optimum classifier is the maximum *a posteriori* probability (MAP) detector, which can then be represented as

$$\hat{\zeta}_i = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \Pr(\zeta_i = k|\boldsymbol{X}). \tag{11}$$

The direct calculation of the posterior probability is numerically challenging due to the high dimension of the variables and parameters. Two unsupervised learning methods, MCMC and EM, are adopted by this article to obtain the classification results.

A. Markov Chain Monte Carlo

The posterior probability $\Pr(\zeta_i = k|\boldsymbol{X})$ can be numerically estimated by using MCMC with Gibbs sampling. Gibbs sampling iteratively takes Monte Carlo samples based on the full conditional distributions of all variables in the mixture model [24]. The samples can be used to obtain an estimate of the posterior probability.

Before starting the iterative process of Gibbs sampling, we need to initialize all the variables within the model, including $\boldsymbol{q}$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_N]$. All variables are first initialized by applying K-means classification on the data. Denote the results of K-means classification as $\zeta_i^{(0)} = k$. Define $\mathcal{S}_k^{(0)} = \{i : \zeta_i^{(0)} = k\}$ as the set of pixels classified into the kth category, and $n_k^{(0)} = |\mathcal{S}_k^{(0)}|$ is the cardinality of $\mathcal{S}_k^{(0)}$. The initial values of the variables can then be calculated as

$$q_k^{(0)} = \frac{n_k^{(0)}}{N}, \quad k = 1, \ldots, K$$

$$\boldsymbol{\mu}_k^{(0)} = \frac{1}{n_k^{(0)}} \sum_{i \in \mathcal{S}_k^{(0)}} \boldsymbol{x}_i, \quad k = 1, \ldots, K$$

$$\boldsymbol{\Sigma}_k^{(0)} = \frac{1}{n_k^{(0)} - 1} \times \sum_{i \in \mathcal{S}_k^{(0)}} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(0)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(0)}\right)^T, \quad k = 1, \ldots, K.$$

Under the Bayesian setting, the unknown parameters are random with prior distributions

$$\pi(q_k) = \text{Dir}(\alpha_k)$$
$$\pi(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_{0k}, \boldsymbol{\Sigma}_{0\,k})$$
$$\pi(\boldsymbol{\Sigma}_k) = \text{InvWish}_p(\boldsymbol{\Psi}, \nu)$$

where Dir and InvWish represent the Dirichlet and inverse Wishart distributions, respectively; α_k, $\boldsymbol{\mu}_{0\,k}$, $\boldsymbol{\Sigma}_{0k}$, $\boldsymbol{\Psi}$, and ν are the hyperparameters of the distributions. Since there is no prior knowledge about these distributions, we assume that $\boldsymbol{\mu}_{0k} = \mathbf{0}_{L'}$, $\boldsymbol{\Sigma}_{0k} = \mathrm{I}_{\boldsymbol{L'}}$, $\boldsymbol{\Psi} = \mathrm{I}_{\boldsymbol{L'}}$, and $\nu = L' + 1$ [25], where L' corresponds to L and $2L$ for the amplitude and complex spectrum analysis, respectively.

Given these priors, the posterior full conditional distributions of these variables can be calculated as follows [19].

1) Posterior distribution of $\boldsymbol{q}$

$$q_k \sim \text{Dir}(\alpha_k + n_k) \tag{12}$$

where n_k is the number of pixels classified into the kth category in the previous iteration.

2) Posterior distribution of $\boldsymbol{\Sigma}_k$

$$\boldsymbol{\Sigma}_k \sim \text{InvWish}_p(\boldsymbol{S} + \boldsymbol{\Psi}, n_k + \nu) \tag{13}$$

where $S = \sum_{i \in \mathcal{S}_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T$, and $\mathcal{S}_k$ is the set of pixels classified into the kth category in the previous iteration.

3) Posterior distribution of $\boldsymbol{\mu}_k$

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \tag{14}$$

Where $\boldsymbol{\Sigma}_p = (\boldsymbol{\Sigma}_{0k}^{-1} + n_k \boldsymbol{\Sigma}_k^{-1})^{-1}$ and $\boldsymbol{\mu}_p = (\boldsymbol{\Sigma}_{0k}^{-1} + n_k \boldsymbol{\Sigma}_k^{-1})^{-1} (\boldsymbol{\Sigma}_{0k}^{-1} \boldsymbol{\mu}_{0k} + \boldsymbol{\Sigma}_k^{-1} \sum_{i \in S_k} \boldsymbol{x}_i)$.

4) Posterior distribution of ζ_i

$$\Pr\left(\zeta_i = k \middle| \boldsymbol{x}_i, [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, q_k]_{k=1}^K\right) = \frac{q_k g(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \zeta_i = k)}{\sum_{p=1}^K q_p g(\boldsymbol{x}_i|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p, \zeta_i = p)}. \tag{15}$$

The Monte Carlo samples of all variables can be iteratively drawn from the above full conditional distributions. The samples are used to numerically approximate the posterior distribution of ζ_i as

$$\Pr(\zeta_i = k|\boldsymbol{X}) = \lim_{M \to \infty} \frac{1}{M} \sum_{it=1}^{M} \mathcal{I}(\zeta_i^{(it)} = k) \tag{16}$$

where $\mathcal{I}(a) = 1$ if $a =$ TRUE and 0 otherwise. MAP detection can then be applied with (16) to perform classification. It should be noted that, before applying the MCMC algorithm, the data vector X might need to be scaled up to avoid numerical underflow during the iteration process. The scaling factor depends on the data values and the precision of the floating number representation used in the computer. In this article, the vectors X are scaled by a factor of 15 before applying the amplitude MCMC algorithm to fresh samples.

B. Expectation Maximization

The posterior distribution of the latent variable $\boldsymbol{\zeta}$ can be alternatively estimated with the EM approach. In this method, we iteratively determine the estimators of the parameters involved in the GMM, $\boldsymbol{\theta} = [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, q_k]_{k=1}^K$, that maximize its log-likelihood function, $\ell(\boldsymbol{\theta}) = \log p(\boldsymbol{X}|\boldsymbol{\theta})$, as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} q_k g(\boldsymbol{x}_i|\zeta_i = k, \boldsymbol{\theta}) \right). \tag{17}$$

It is difficult to directly maximize the log-likelihood function $\ell(\boldsymbol{\theta})$ due to the logarithm of summation. The EM algorithm iteratively maximizes the log-likelihood function by employing an expectation step (E-step) and maximization step (M-step) [26].

1) E-Step: In the E-step of the mth iteration, the algorithm first calculates the posterior probability of ζ_i by using (15), and the result is denoted as

$$\gamma_{ik}^{(m)} = \Pr\left(\zeta_i = k \middle| \boldsymbol{x}_i, \boldsymbol{\theta}^{(m)}\right) \tag{18}$$

where $\boldsymbol{\theta}^{(m)} = [\boldsymbol{\mu}_k^{(m)}, \boldsymbol{\Sigma}_k^{(m)}, q_k^{(m)}]_{k=1}^K$ are the model parameters from the mth iteration.

2) M-Step: In the M-step, the algorithm maximizes the conditional expectation of the joint log-likelihood function of $\boldsymbol{y}$ and ζ_i, which can be expressed as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{N} \mathbb{E}_{\zeta_i|\boldsymbol{\theta}^{(m)}}\left[\log p\left(\boldsymbol{x}_i, \zeta_i \middle| \boldsymbol{\theta}\right)\right] \tag{19}$$

where the expectation is performed with respect to the posterior distribution of $\Pr(\zeta_i = k|\boldsymbol{\theta}^{(m)})$.

Calculating the conditional expectation in (19) yields

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = &\sum_{k=1}^{K} \eta_k^{(m)} \left[\log q_k - \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_k|\right] \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik}^{(m)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \end{aligned} \tag{20}$$

where

$$\eta_k^{(m)} = \sum_{i=1}^{N} \gamma_{ik}^{(m)}. \tag{21}$$

Maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ with respect to $\boldsymbol{\theta}$ yields the following parameter estimators.

1) Estimator of q_k

$$q_k^{(m+1)} = \frac{\eta_k^{(m)}}{N}. \tag{22}$$

2) Estimator of $\boldsymbol{\mu}_k$

$$\boldsymbol{\mu}_k^{(m+1)} = \frac{1}{\eta_k^{(m)}} \sum_{i=1}^{N} \gamma_{ik}^{(m)} \boldsymbol{x}_i. \tag{23}$$

3) Estimator of $\boldsymbol{\Sigma}_k$

$$\boldsymbol{\Sigma}_k^{(m+1)} = \frac{1}{\eta_k} \sum_{i=1}^{} \gamma_{ik}^{(m)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(m+1)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(m+1)}\right)^T. \tag{24}$$

The convergence of the algorithm is guaranteed because the M-step will always increase the log-likelihood function $\ell(\boldsymbol{\theta})$ [26].

VI. Experimental Results

The performance of the newly proposed LOOP algorithm with unsupervised learning is quantitatively evaluated in this section with THz images of freshly excised breast cancer tissue. All the source codes used for this analysis are available in [27].

The classification results from THz images of freshly excised tissues are compared to histopathology results from the corresponding FFPE tissues. Since the FFPE samples are obtained by fixing fresh tissue samples in paraffin, there is usually a significant mismatch between the shapes of the FFPE and fresh tissues. Thus, a direct pixel-by-pixel comparison between the results from the THz image and the histopathology results is not possible.

To enable quantitative evaluations of the results, we employ the image morphing algorithm [5] on the pathology results to create a reference image with the same size and resolution as the THz image. The morphed pathology image is used to represent the real classification of each pixel according to the pathology report. Such a morphing method enables the quantitative evaluation of the detection results through pixel-by-pixel comparisons between the detection results and the morphed pathology results. This comparison is summarized in a receiver operating characteristic (ROC) curve, which is a plot showing the true detection rate as a function of the false detection rate. Since the results of the statistical analysis are represented as the probability of each pixel belonging to different regions, we can adjust the probability threshold for the detection of a certain region to obtain different points on the ROC curve.

In the proposed LOOP algorithm, each pixel is summarized as a low-dimensional vector extracted from the THz spectrum. During the analysis, the LOOP algorithm was applied to both the amplitude spectrum and the complex spectrum of the THz image, respectively. For each tissue sample, results from various sizes of the low-dimensional vectors obtained from the LOOP algorithm are compared, and the one that yields the best performance is presented. In addition, we will compare the performance of the LOOP algorithm with several existing algorithms, including the 1-D MCMC algorithm that summarizes each pixel into a 1-D scalar [5], [8], and the PCA algorithm [28]. It is important to mention that the 1-D MCMC algorithm classifies the regions according to the spectral power of the frequency-domain signal per pixel for fresh tissue, and the peak reflection of the time-domain signal for block tissue, respectively [8]. All detection algorithms are applied to three different human breast tumor samples, and the corresponding results are given in this section.

A. Results From Freshly Excised Samples

We first present the results obtained by analyzing three human breast cancer tissue samples: ND10898, ND15526, and ND15588, with dimensions 15×15 mm, 8.7×13 mm, and 8×15.3 mm, respectively. We receive fresh tissue of thickness ranging from 3 to 4 mm. As reported in [7], the tissue have high-absorption coefficient ranging from $\sim$100 to 700 cm^{-1} in the frequency range of 0.1–3.5 THz. Thus, the multiple reflection interference inside the tissue becomes insignificant. For example, at 0.5 THz, the signal penetration depth is $\sim$276 μm in cancer [29]; therefore, the reflected signal from tissue of less than $\sim$2-mm thickness could be adversely affected by the

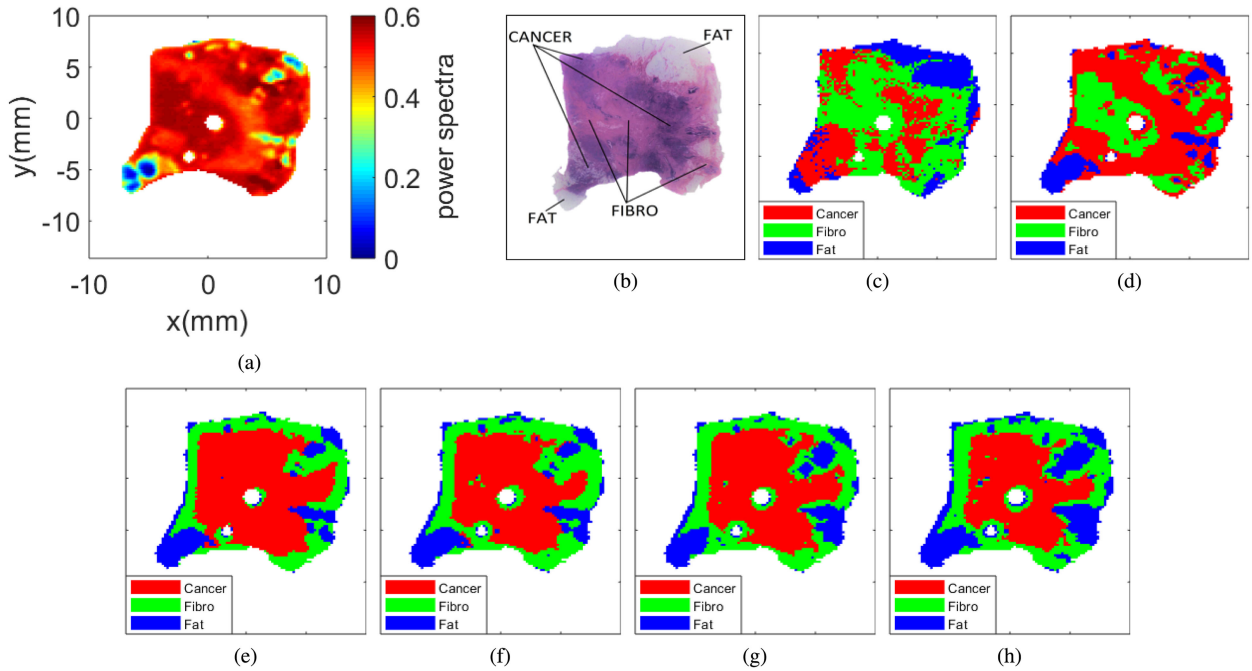

Fig. 3. Sample ND10898 fresh. (a) THz image. (b) Pathology image. (c) Morphed pathology. (d) 1D MCMC model. (e) 2-D amplitude MCMC model. (f) 2-D amplitude EM model. (g) 4-D complex MCMC model. (h) 4-D complex EM model.

multiple reflection. These samples contain three regions: cancer, collagen or fibro, and fat.

Fig. 3(a) shows the THz image collected from sample ND10898 while it was still fresh, where each pixel represents the power spectra of its THz waveform [8]. Fig. 3(b) represents the histopathology results obtained by analyzing the FFPE tissue sample fixed in paraffin, which corresponds to the gold standard within cancer detection. Fig. 3(c) shows the morphed pathology mask obtained by employing the morphing algorithm [5]. The morphed pathology mask is used as a benchmark for the THz image classification results. The white spots within all the images in Fig. 3 represent air bubbles (artifact from the data collection process) that were removed before further processing to avoid data contamination.

The classification results of the THz image obtained by using the 1-D MCMC approach [5], [8], 4-D MCMC with amplitude spectrum, 2-D EM with amplitude spectrum, 4-D MCMC with complex spectrum, and 4-D EM with complex spectrum are presented in Fig. 3(d)–(h), respectively. The 2-D and 4-D results are obtained by using the proposed LOOP algorithm. By visually inspecting the classification models results side-by-side, we can observe that the fibro detection in the 1-D MCMC approach is the best among all the models at the cost of a large misclassification of cancer. On the other hand, the correlation among the cancer and fat regions is improved in the 2-D and 4-D models presented in Fig. 3(e)–(h) when compared to the morphed pathology results in Fig. 3(c).

To quantify the performance of each model, the corresponding ROC curves of the classification results of sample ND10898 fresh are presented in Fig. 4. The ROC curves are obtained by performing pixel-by-pixel comparisons between the detection results and the morphed pathology results. The areas underneath the ROC curves are listed in Table I. All results obtained with the proposed LOOP algorithm perform significantly better than the 1-D MCMC approach [8] for both cancer and fat, while the detection of fibro is better in 1-D MCMC. The results from 2-D feature vectors achieve larger cancer ROC areas (∼60%) than those from the 1-D approach (∼50%). Hence, we can state that the analysis of higher dimensional feature vectors significantly improves the detection accuracy. In terms of areas underneath the ROC curves, 2-D amplitude EM achieves the best performance for cancer and fat detection.

Similarly, Fig. 5(a) represents the THz image collected from sample ND15526 fresh. Fig. 5(b) and (c) correspond to the original and morphed histopathology results. Fig. 5(d)–(h) shows the classification results for 1-D MCMC, 2-D amplitude MCMC, 2-D amplitude EM, 3-D complex MCMC, and 6-D complex EM, respectively. Visually, 1-D MCMC, 2-D amplitude MCMC, and 3-D complex MCMC present similar classification areas with good cancer correlation, but with poor collagen detection. On the contrary, 2-D amplitude EM and 6-D complex EM present a better collagen detection at the cost of large cancer regions misclassification.

Fig. 6 presents the ROC curves for sample ND15526 fresh and their areas under the ROC curves are presented in Table I. We can observe that the detection of cancer and fat is comparable among the 1D MCMC approach and most of the higher dimensional models, with 1-D MCMC being slightly better. Overall the best classification results are obtained by the 6-D complex EM approach. This method achieved areas under the ROC of 77% or above for all the regions presented in this sample.

Fig. 7(a) shows the THz image collected from sample ND15588 while it was still fresh. Fig. 7(b) represents the

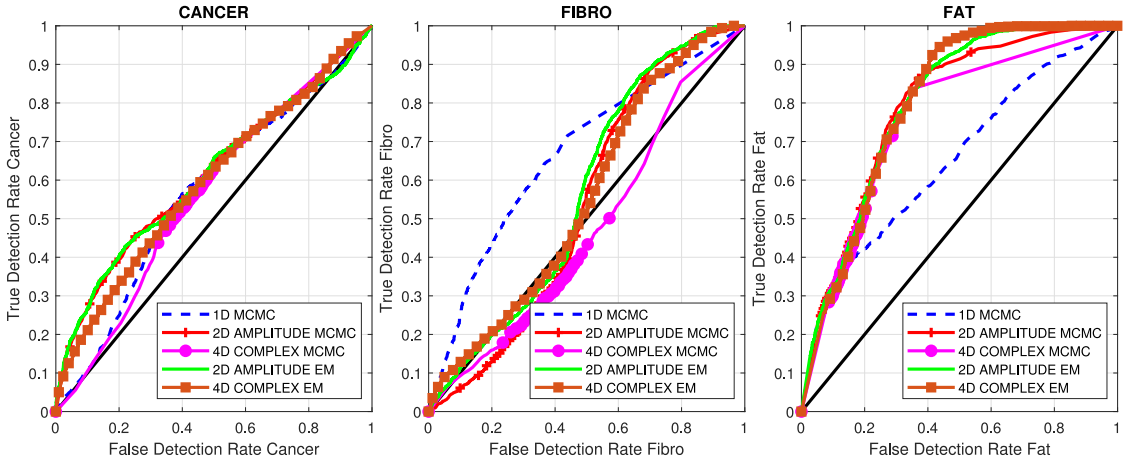

Fig. 4. ROC curves for sample ND10898 fresh.

TABLE I
AREAS UNDER THE ROC CURVES

ND10898 Fresh (15 × 15 mm)					
Region	1D MCMC	2D amplitude MCMC	2D amplitude EM	4D complex MCMC	4D complex EM
Cancer	0.5676	0.6130	0.6101	0.5631	0.5934
Fibro	0.6682	0.5370	0.5663	0.4723	0.5528
Fat	0.6530	0.7885	0.7963	0.7571	0.7971
ND15526 Fresh (8.7 × 14 mm)					
Region	1D MCMC	2D amplitude MCMC	2D amplitude EM	3D complex MCMC	6D complex EM
Cancer	0.7468	0.7122	0.7353	0.7011	0.7750
Collagen	0.6458	0.5576	0.6027	0.6008	0.7705
Fat	0.8390	0.8215	0.8247	0.8256	0.8327
ND15588 Fresh (8 × 15.3 mm)					
Region	1D MCMC	2D amplitude MCMC	2D amplitude EM	3D complex MCMC	4D complex EM
Cancer	0.6338	0.7435	0.7469	0.7481	0.7083
Collagen	0.6521	0.7338	0.7412	0.7286	0.7451
Fat	0.7372	0.7619	0.7685	0.7941	0.7759
ND15588 Block (7.5 × 14.9 mm)					
Region	1D MCMC	6D amplitude MCMC	6D amplitude EM	2D complex MCMC	2D complex EM
Cancer	0.7305	0.7997	0.7977	0.6735	0.6752
Collagen	0.4843	0.6366	0.6280	0.6052	0.6668
Fat	0.8743	0.7999	0.7674	0.7109	0.7588

histopathology results obtained by analyzing the corresponding FFPE tissue sample fixed in paraffin. Fig. 7(c) shows the morphed pathology mask obtained by employing the morphing algorithm [5]. The classification results obtained by using the 1-D MCMC approach, 4-D MCMC with amplitude spectrum, 2-D EM with amplitude spectrum, 4-D MCMC with complex spectrum, and 4-D EM with complex spectrum are presented in Fig. 7(d)–(h), respectively. For the 1-D MCMC approach, large portions of the cancer regions are misclassified as collagen. For the 2-D and 4-D results obtained with amplitude spectrum, there is a slight improvement in the detection of the cancer region for both MCMC and EM algorithms. For the high-dimensional results obtained from the complex spectrum, the detection of cancer and fat slightly improves when compared to their amplitude counterparts, but at the cost of a higher misclassification of collagen. It is important to mention that the surrounding cancer zones in Fig. 7(e)–(h) that do not correlate with the histopathology results correspond to the misclassification of the detection algorithms.

The corresponding ROC curves of the classification results of sample ND15588 fresh are presented in Fig. 8. The areas underneath the ROC curves are listed in Table I. All results obtained with the proposed LOOP algorithm perform considerably better than the 1-D MCMC approach [8]. The results from 2-D, 3-D, and 4-D feature vectors achieve larger ROC areas ($\sim 70\%$) than those from the 1-D approach ($\sim 60\%$). Thus, we can conclude that increasing the dimension of the feature vector by just one dimension over the 1-D approach can achieve apparent performance improvement. In terms of areas underneath the ROC curves, 2-D amplitude EM achieves the best performance for all the regions.

B. Results From FFPE Block Sample

We also analyze the classification results obtained by using the THz image of FFPE block sample, where the image is obtained by scanning the paraffin embedded block sample. The THz image of sample ND15588 block is shown in Fig. 9(a),

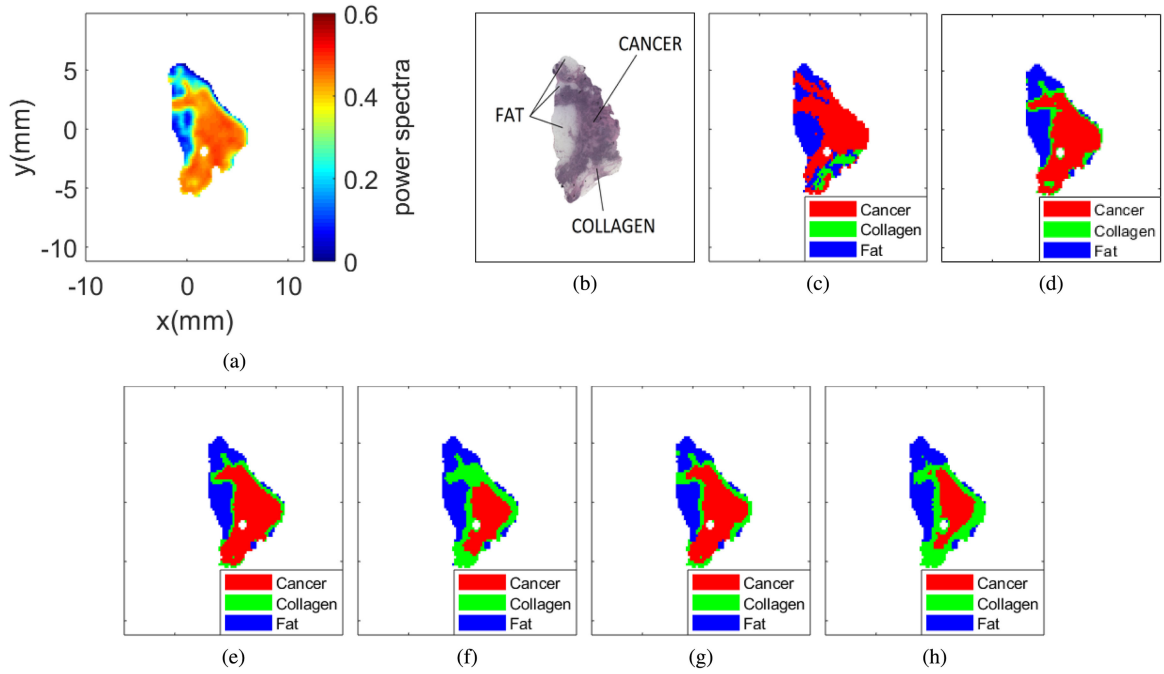

Fig. 5. Sample ND15526 fresh. (a) THz image. (b) Pathology image. (c) Morphed pathology. (d) 1-D MCMC model. (e) 2-D amplitude MCMC model. (f) 2-D amplitude EM model. (g) 3-D complex MCMC model. (h) 6-D complex EM model.

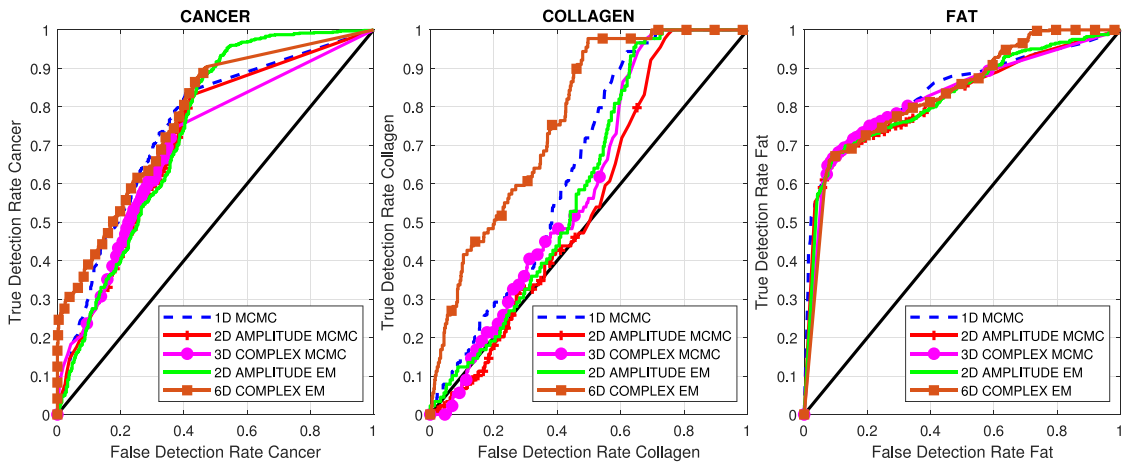

Fig. 6. ROC curves for sample ND15526 fresh.

where each pixel is represented by using the peak reflection of the THz waveform [8]. The dimensions of this block sample are 7.5 × 14.9 mm and its thickness is ∼3–4 mm. As explained in [8], the block tissue is sensitive to multiple reflections in the frequency domain due to its low absorbance; hence, the power spectra are not utilized for this type of samples in the 1-D case. For imaging the dehydrated tissue block (FFPE), the time-domain peak reflection from each pixel on the surface is measured. These peaks are not affected by the multiple reflections due to the difference in arrival times. Even though this set of results corresponds to the same sample, as presented in Fig. 7(a), this image was collected from scanning the paraffin block tissue obtained after the pathology process. As a result, the THz image of FFPE block tissue is different from that of its fresh counterpart shown in Fig. 7(a). It is important to mention that we include the results obtained from block tumor samples to illustrate the behavior of the algorithms within this sample type. Since the region detection among block samples is of limited clinical interests, we present one sample only for this purpose.

The corresponding histopathology results and morphed histopathology mask are shown in Fig. 9(b) and (c), respectively. The classification results obtained by using the 1-D MCMC approach [5], [8], 6-D MCMC with amplitude spectrum, 6-D EM with amplitude spectrum, 2-D MCMC with complex spectrum, and 2-D EM with complex spectrum are presented in Fig. 9(d)–(h), respectively. The corresponding ROC curves are

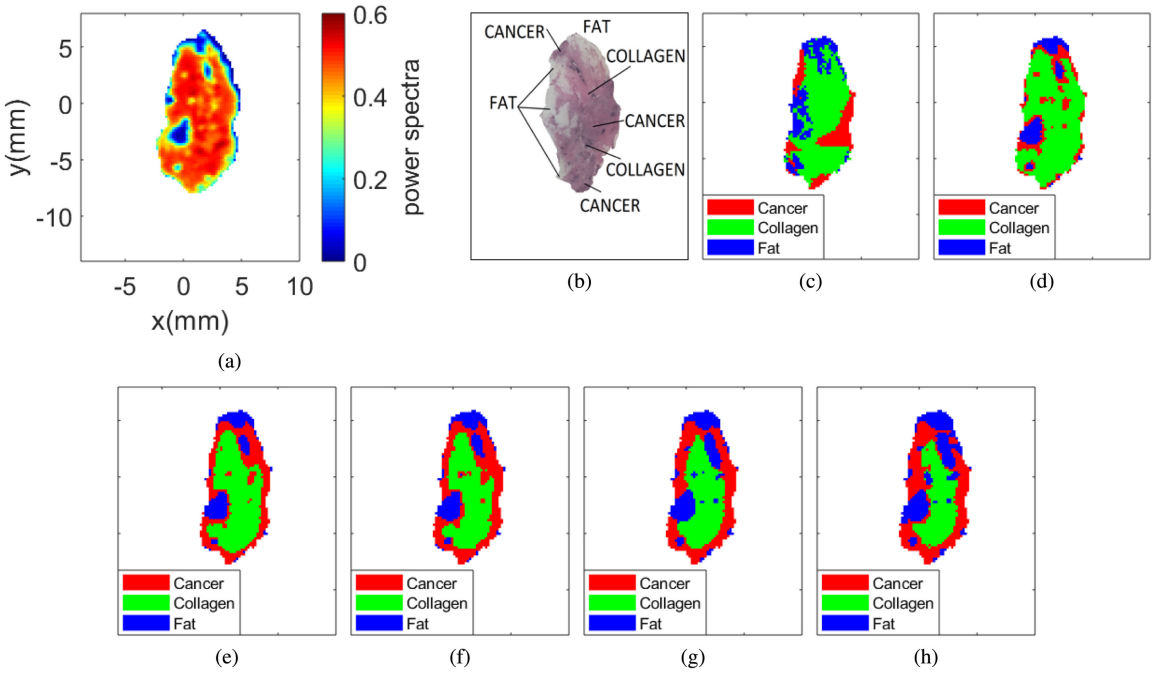

Fig. 7. Sample ND15588 fresh. (a) THz image. (b) Pathology image. (c) Morphed pathology. (d) 1-D MCMC model. (e) 2-D amplitude MCMC model. (f) 2-D amplitude EM model. (g) 3-D complex MCMC model. (h) 4-D complex EM model.

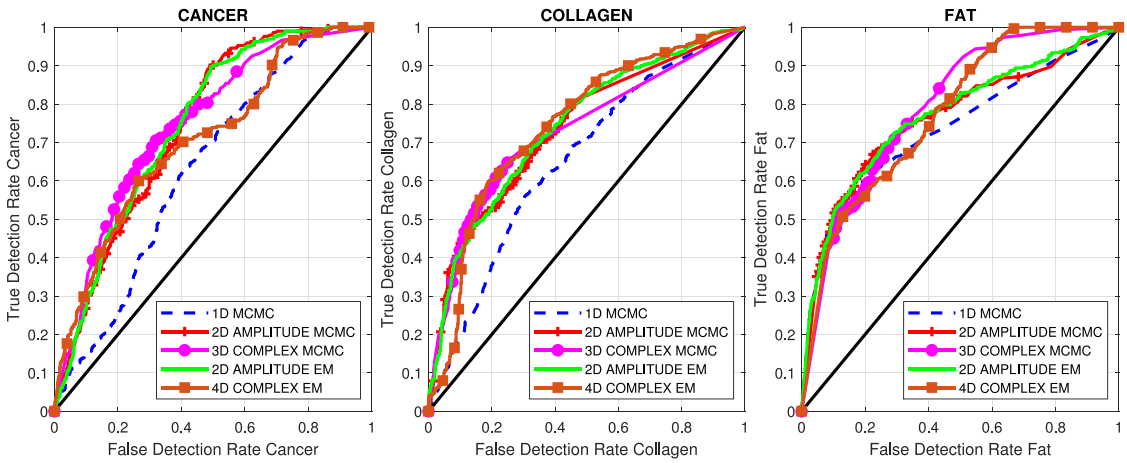

Fig. 8. ROC curves for sample ND15588 fresh.

given in Fig. 10. The areas underneath the ROC curves are listed in Table I.

Visually, the results obtained from the 6-D amplitude MCMC and 6-D amplitude EM models have the best overall correlation with the histopathology results. This is corroborated by the ROC curves for the cancer region. The cancer ROC areas of the 6-D amplitude MCMC and EM approaches are 79.97% and 79.77%, which are significantly higher than other methods with ROC areas ranging from 67.35% to 73.05%. It should be noted that the relatively large cancer ROC area of the 1-D MCMC model is achieved at the cost of extremely poor performance of collagen, where the majority of the collagen pixels are misclassified as cancer, as shown in Fig. 9(d). In terms of the collagen ROC area, the 6-D amplitude MCMC and 2-D complex EM models achieve the best performance among all cases, with that of 2-D complex EM being better. However, the 2-D complex EM model has a large misclassification of cancer. For the fat region, the 1-D MCMC and 6-D amplitude MCMC models have the best performance, followed by the 6-D amplitude EM model. The 6-D amplitude MCMC model has the best overall performance in terms of visual correlation and ROC areas, which are comparable to the results obtained from the 6-D amplitude EM model. The ROC areas of the 6-D amplitude MCMC model are 79.97%, 63.66%, and 79.99%, respectively.

It is to be noted that heterogeneous human tissues have an uneven surface. This necessitates some "facing in" of the paraffin

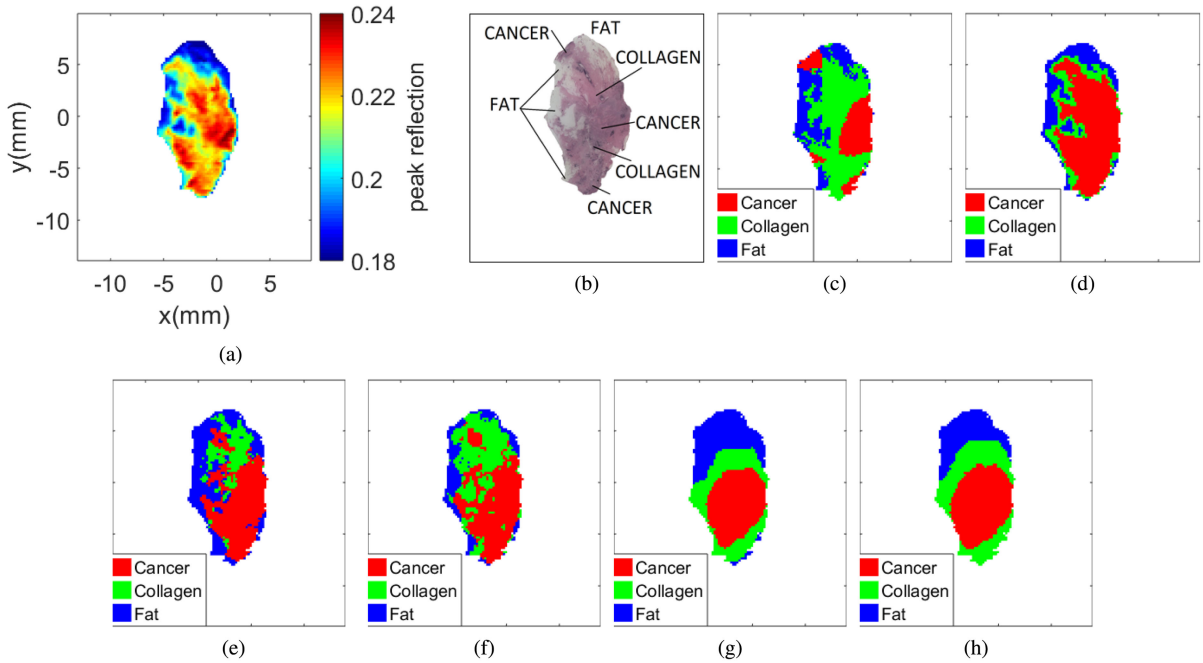

Fig. 9. Sample ND15588 block. (a) THz image. (b) Pathology image. (c) Morphed pathology. (d) 1-D MCMC model. (e) 6-D amplitude MCMC model. (f) 6-D amplitude EM model. (g) 2-D complex MCMC model. (h) 2-D complex EM model.

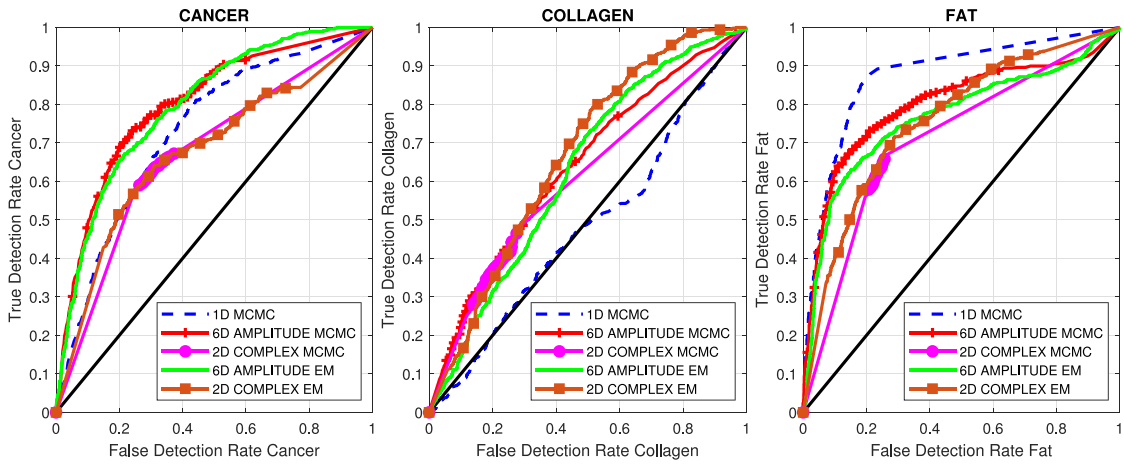

Fig. 10. ROC curves for sample ND15588 block.

block in order to obtain a full/intact tissue section. In general, "facing in" the block will result in the loss of approximately 100 μm off the uneven surface. Therefore, THz imaging of dehydrated samples, such as FFPE, has shown better correlation with pathology because both images were taken from the same surface. On the other hand, the THz imaging of fresh samples was taken from different surfaces. Furthermore, the contrast between cancer and healthy nonfatty tissue is affected by the water content in both.

Overall, the amplitude-based models perform better than the complex spectrum models for block tissues. Visually, the results obtained with complex spectrum do not correlate well with the morphed pathology results. Hence, utilizing both amplitude and phase information of the THz spectrum might negatively impact the overall classification results with FFPE tissue samples.

C. Comparison With Other Methods

The performance of the proposed LOOP algorithm with unsupervised statistical learning is compared to several other commonly used algorithms in the literature, including PCA [28], K-means [30], and SVM [31]. PCA is a widely used dimension reduction algorithm. K-means and SVM are commonly used unsupervised and supervised machine learning algorithms, respectively. The comparison is performed by using sample ND15588 fresh. For fairness of comparison, the same dimension is used

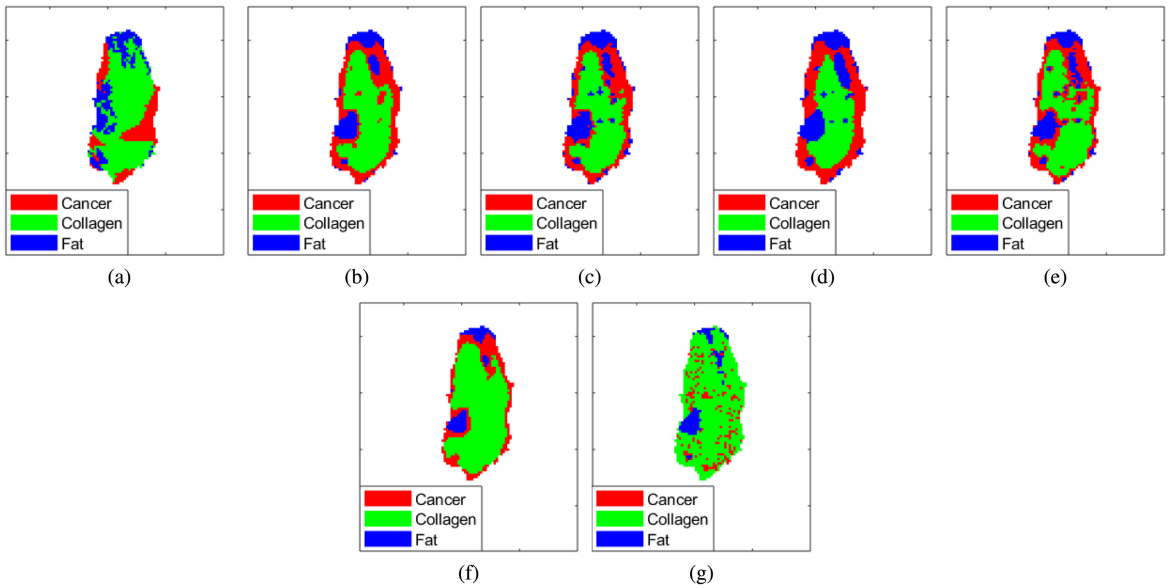

Fig. 11. Sample ND15588 fresh with different classification methods. (a) Morphed histopathology mask. (b) 2-D amplitude MCMC with LOOP. (c) 2-D amplitude MCMC with PCA. (d) 3-D complex MCMC with LOOP. (e) 3-D complex MCMC with PCA. (f) K-means clustering with full spectrum. (g) SVM clustering with full spectrum.

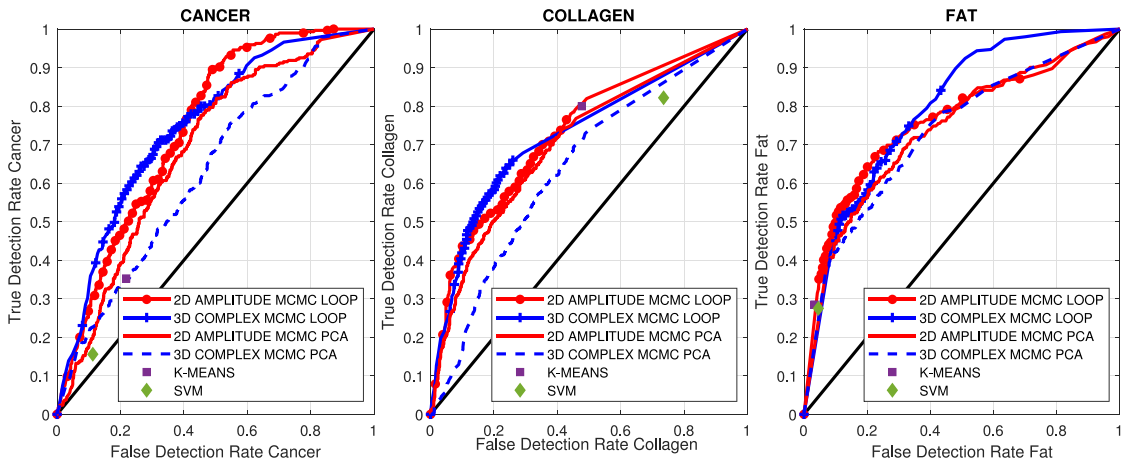

Fig. 12. ROC curves for sample ND15588 fresh with different classification methods.

TABLE II
AREAS UNDER THE ROC CURVES FOR SAMPLE ND15588 FRESH: LOOP VERSUS PCA

Region	2D amplitude MCMC with LOOP	2D amplitude MCMC with PCA	3D complex MCMC with LOOP	3D complex MCMC with PCA
Cancer	0.7435	0.6871	0.7481	0.6307
Collagen	0.7338	0.7067	0.7286	0.6418
Fat	0.7619	0.7387	0.7941	0.7327

by both PCA and LOOP. The low-dimensional vectors obtained from PCA or LOOP are further processed by using amplitude or complex MCMC. No dimension reduction is applied to either K-means or SVM. Since SVM is a supervised algorithm, it is first trained with sample ND15526, and the trained model was then applied to sample ND15588. The morphed histopathology results are presented in Fig. 11(a). The classification results of 2-D amplitude MCMC with LOOP, 2-D amplitude MCMC with PCA, 3-D complex MCMC with LOOP, 3-D complex MCMC with PCA, K-means, and SVM are shown in Fig. 11(b)–(g), respectively. The corresponding ROC curves of the cancer, collagen, and fat regions are shown in Fig. 12. The areas underneath the ROC curves are listed in Table II. It should be noted that K-means and SVM are hard-clustering techniques; therefore, the results of K-means or SVM are fixed, and they cannot be tuned based on the tradeoff between the true positive and false

TABLE III
DETECTION RATES FOR SAMPLE ND15588 FRESH: K-MEANS AND SVM

Region	K-Means		SVM	
	True Detection Rate	False Detection Rate	True Detection Rate	False Detection Rate
Cancer	0.3525	0.2179	0.1559	0.1127
Collagen	0.7998	0.4774	0.8217	0.7353
Fat	0.2848	0.0331	0.2748	0.0455

positive probabilities. Consequently, the results from K-means and SVM are represented as single dots on the ROC curves in Fig. 12. The true and false detection rates of K-Means and SVM for all the regions within this sample are shown in Table III.

Overall, the 2-D amplitude MCMC and the 3-D complex MCMC models with LOOP achieves the best performance among all the different methods. K-means achieves comparable results with respect to 2-D amplitude MCMC for both collagen and fat, but its detection of cancer is much worse than MCMC. The SVM method shows poor detection of cancer and a large missclassification of collagen. The LOOP algorithm outperforms the PCA algorithm in all three regions within the tumor sample. The areas under the ROC curves for the PCA approaches achieve values of 63.07%–73.87% for all regions, while the LOOP counterparts achieve areas of 72.86%–79.41%. Thus, the proposed LOOP algorithm can achieve better performance than the well-established algorithms, such as PCA, K-means, and SVM.

VII. CONCLUSION

A new dimension reduction algorithm has been proposed to extract the salient information embedded in THz images of cancer tissues. The LOOP algorithm summarizes the wide spectrum of each pixel in the THz image as a low-dimensional feature vector, which is then modeled by using multivariate GMMs. The low-dimensional feature vectors were utilized by MCMC or EM algorithms to classify the different regions within a sample tissue. The newly proposed algorithm was applied to human breast cancer tissue samples with three regions. Experiment results have demonstrated that the LOOP method achieves apparent performance improvement over existing approaches, such as the 1-D MCMC approach [5], [8]. For example, the areas under the cancer ROC curves have been improved from 63.38% to 74.69% by simply replacing the 1-D features in the 1-D MCMC algorithm with 2-D feature vectors extracted from the LOOP algorithm in sample ND15588 fresh.

In general, the EM algorithm with the LOOP method achieves the best overall performance, for both freshly excised tissues and FFPE block tissues. In particular, the algorithms present promising results for freshly excised human tissues with at least 60%–70% of areas underneath the ROC curves. This represents an important milestone in the region classification of human breast cancer tissues, which are significantly more heterogeneous and complex than the xenograft mice tissues used in [5] and [8]. The classification of tumor tissues with three or more regions still remains as a significant challenge for future works. To further improve the classification performance, we plan to explore the spatial correlation among neighboring pixels by using graph theory and spatial statistics, which can identify and model the dependence among pixels in a given neighborhood.

ACKNOWLEDGMENT

The authors would like to thank the Oklahoma Animal Disease Diagnostic Laboratory and Oklahoma State University for their support in handling the tumor samples presented in this article.

REFERENCES

[1] World Cancer Research Fund, "Breast cancer statistics," 2018. Accessed: May 29, 2019. [Online]. Available: https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statisti cs
[2] Breastcancer.org, "What is lumpectomy?" Mar. 4, 2015. Accessed: May 29, 2019. [Online]. Available: https://www.breastcancer.org/treatment/surgery/lumpectomy/what_is
[3] L. Havel, H. Naik, L. Ramirez, M. Morrow, and J. Landercasper, "Impact of the SSO-ASTRO margin guideline on rates of re-excision after lumpectomy for breast cancer: A meta-analysis," *Ann. Surg. Oncol.*, vol. 26, no. 5, pp. 1238–1244, May 2019.
[4] B. C. Q. Truong, A. J. Fitzgerald, S. Fan, and V. P. Wallace, "Concentration analysis of breast tissue phantoms with terahertz spectroscopy," *Biomed. Opt. Express*, vol. 9, no. 3, pp. 1334–1349, Mar. 2018.
[5] T. Chavez, T. Bowman, J. Wu, K. Bailey, and M. El-Shenawee, "Assessment of terahertz imaging for excised breast cancer tumors with image morphing," *J. Infrared, Millimeter, Terahertz Waves*, vol. 39, no. 12, pp. 1283–1302, Dec. 2018.
[6] N. Vohra *et al.*, "Pulsed terahertz reflection imaging of tumors in a spontaneous model of breast cancer," *Biomed. Phys. Eng. Express*, vol. 4, no. 6, Oct. 2018, Art. no. 065025.
[7] T. Bowman, N. Vohra, K. Bailey, and M. O. El-Shenawee, "Terahertz tomographic imaging of freshly excised human breast tissues," *J. Med. Imag.*, vol. 6, no. 2, 2019, Art. no. 023501.
[8] T. Bowman *et al.*, "Pulsed terahertz imaging of breast cancer in freshly excised murine tumors," *J. Biomed. Opt.*, vol. 23, no. 2, 2018, Art. no. 026004.
[9] J. Shi *et al.*, "Automatic evaluation of traumatic brain injury based on terahertz imaging with machine learning," *Opt. Express*, vol. 26, no. 5, pp. 6371–6381, Mar. 2018.
[10] Y. Cao *et al.*, "Inspecting human colon adenocarcinoma cell lines by using terahertz time-domain reflection spectroscopy," *Spectrochimica Acta Part A: Molecular Biomolecular Spectroscopy*, vol. 211, pp. 356–362, 2019.
[11] Y. V. Kistenev, A. V. Borisov, M. A. Titarenko, O. D. Baydik, and A. V. Shapovalov, "Diagnosis of oral lichen planus from analysis of saliva samples using terahertz time-domain spectroscopy and chemometrics," *J. Biomed. Opt.*, vol. 23, no. 4, 2018, Art. no. 045001.
[12] H. Liu *et al.*, "Dimensionality reduction for identification of hepatic tumor samples based on terahertz time-domain spectroscopy," *IEEE Trans. THz Sci. Technol.*, vol. 8, no. 3, pp. 271–277, May 2018.
[13] Z. Zhang, H. Liu, and C. Zhang, "Terahertz pulse data dimensional reduction and classification for hepatic tissue samples," in *Proc. 43rd Int. Conf. Infrared, Millimeter, Terahertz Waves*, Sep. 2018, pp. 1–3.
[14] M. W. Ayech and D. Ziou, "Segmentation of terahertz imaging using k-means clustering based on ranked set sampling," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 2959–2974, 2015.
[15] X. Yin, W. Mo, Q. Wang, and B. Qin, "A terahertz spectroscopy nondestructive identification method for rubber based on CS-SVM," *Adv. Condensed Matter Phys.*, vol. 2018, 2018, Art. no. 1618750.
[16] Y. Li, X. A. Shen, R. L. Ewing, and J. Li, "Terahertz spectroscopic material identification using approximate entropy and deep neural network," in *Proc. IEEE Nat. Aerosp. Electron. Conf.*, Jun. 2017, pp. 52–56.

[17] T. Chavez, T. Bowman, J. Wu, M. El-Shenawee, and K. Bailey, "Cancer classification of freshly excised murine tumors with ordered orthogonal projection," in *Proc. IEEE Int. Symp. Antennas Propag./USNC-URSI Radio Sci. Meeting*, Jul. 2019, pp. 525–526.
[18] D. Reynolds, *Gaussian Mixture Models*. Boston, MA, USA: Springer, 2015, pp. 827–832.
[19] X. Zhang, J. Bolton, and P. Gader, "A new learning method for continuous hidden Markov models for subsurface landmine detection in ground penetrating radar," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 3, pp. 813–819, Mar. 2014.
[20] H. D. Vargas Cardona, Á. Á. Orozco, and M. A. Álvarez, "Unsupervised learning applied in MER and ECG signals through Gaussians mixtures with the expectation-maximization algorithm and variational Bayesian inference," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Jul. 2013, pp. 4326–4329.
[21] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 5, pp. 1087–1095, Sep. 1994.
[22] S. Guha and A. R. Lamichhane, "Document classification after dimension reduction through a modified Gram-Schmidt process," in *Wireless Networks and Computational Intelligence*, K. R. Venugopal and L. M. Patnaik, Eds. Berlin, Germany: Springer, 2012, pp. 236–243.
[23] D. Maiwald and D. Kraus, "Calculation of moments of complex Wishart and complex inverse Wishart distributed matrices," *Proc. Inst. Elect. Eng.—Radar, Sonar Navigat.*, vol. 147, no. 4, pp. 162–168, Aug. 2000.
[24] I. Yildirim, "Bayesian inference: Gibbs sampling," Univ. Rochester, Rochester, NY, USA, Tech. Note, 2012.
[25] I. Alvarez, J. Niemi, and M. Simpson, "Bayesian inference for a covariance matrix," in *Proc. 26th Annu. Conf. Appl. Statist. Agriculture*, Apr. 27–29, 2014, pp. 71–82.
[26] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Springer Series in Statistics), vol. 1. New York, NY, USA: Springer, 2001.
[27] T. Chavez, N. Vohra, J. Wu, K. Bailey, and M. El-Shenawee, Source code for "Breast cancer detection with low-dimension ordered orthogonal projection in terahertz imaging." [Online]. Available: https://github.com/taxe10/LOOP. Accessed on: Oct. 15, 2019.
[28] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Rev.: Comput. Statist.*, vol. 2, no. 4, pp. 433–459, Aug. 2010.
[29] M. El-Shenawee, N. Vohra, T. Bowman, and K. Bailey, "Cancer detection in excised breast tumors using terahertz imaging and spectroscopy," *Biomed. Spectrosc. Imag.*, vol. 8, nos. 1/2, pp. 1–9, Jul. 2019.
[30] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[31] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

Tanny Chavez (S'19) received the B.S. degree in electronics and telecommunications engineering from the Escuela Superior Politecnica del Litoral, Guayaquil, Ecuador, in 2015, and the M.S. degree in electrical engineering in 2018 from the University of Arkansas, Fayetteville, AR, USA, where she is currently working toward the doctoral degree in electrical engineering.

Her research interests include statistical signal processing for the detection of breast cancer in terahertz imaging.

Nagma Vohra (S'18) received the B.S. degree in electronics and communication engineering from Guru Nanak Dev University, Amritsar, India, in 2014, and the M.S. degree in communication engineering from the Vellore. Institute of Technology University, Vellore, India, in 2017. She is currently working toward the Ph.D. degree in electrical engineering with the University of Arkansas, Fayetteville, AR, USA, with the focus on material characterization at microwave and millimeter-wave frequencies.

Jingxian Wu (S'04–M'06–SM'15) received the B.S. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1998, the M.S. degree from Tsinghua University, Beijing, in 2001, and the Ph.D. degree from the University of Missouri, Columbia, MO, USA, in 2005, all in electrical engineering.

He is currently a Professor with the Department of Electrical Engineering, University of Arkansas, Fayetteville, AR, USA. His research interests mainly focus on biomedical signal processing, wireless communications, cybersecurity for smart grids, and statistical data analytics.

Dr. Wu served as a Symposium or Track Co-Chair for a number of international conferences, such as the 2012 and 2019 IEEE International Conference on Communications, the 2009, 2015, and 2017 IEEE Global Telecommunications Conference, etc. He served as an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2007 to 2011 and an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2011 to 2016. He is an Associate Editor for the IEEE ACCESS.

Keith Bailey received the Doctor of Veterinary Medicine and Doctor of Philosophy (Pathobiology Area Program) degrees from the University of Missouri, Columbia, MO, USA, in 1992 and 1996, respectively, and a Board Certification from the American College of Veterinary Pathologists, Madison, WI, USA, in 2000.

He currently serves as a Comparative Pathologist with the University of Illinois at Urbana–Champaign, Urbana, IL, USA. His professional experiences include providing anatomic pathology support during the evaluation of drug candidate molecules for use as human therapeutics.

Magda El-Shenawee (SM'02) received the Ph.D. degree from the University of Nebraska–Lincoln, Lincoln, NE, USA, in 1991.

He is currently a Professor of Electrical Engineering with the University of Arkansas, Fayetteville, AR, USA, since 2001. Her background is in electromagnetics, theory, measurements, and computational techniques. Her educational goals are promoting the online antenna courses for industry and the open electromagnetic laboratory for undergraduate students. She authored or coauthored more than 230 papers in refereed journals and conference proceedings. Her research interests included experimental terahertz imaging and spectroscopy, breast cancer imaging, image reconstruction algorithms, antennas design and measurements, computational electromagnetics, biopotentials, and biomagnetics of breast cancerous cells. Her current research focuses on terahertz imaging and spectroscopy of breast tumor's margins and on material characterization in the microwave, millimeter wave, and terahertz frequency bands.