Biomedical Signal Processing and Control 70 (2021) 102949

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Supervised Bayesian learning for breast cancer detection in terahertz imaging

Tanny Chavez [a], Nagma Vohra [a], Keith Bailey [b], Magda El-Shenawee [a], Jingxian Wu [a,*]

[a] Department of Electrical Engineering, University of Arkansas, Fayetteville, AR 72701 USA
[b] Veterinary Diagnostic Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61802, USA

ARTICLE INFO

Keywords:
Breast cancer
Multinomial probit regression
Random Fourier features
Terahertz imaging

ABSTRACT

This paper proposes a supervised multinomial Bayesian learning algorithm for breast cancer detection using terahertz (THz) imaging of freshly excised murine tumors. The proposed algorithm utilizes a multinomial Bayesian probit regression approach, which establishes the link between THz data and classification results by using two different models, a polynomial regression model and a kernel regression model. Such a model-based learning approach employs only a small number of model parameters, thus it requires much less training data when compared with alternative deep learning methods. The training phase of the algorithm is performed by using the histopathology results of formalin-fixed, paraffin embedded (FFPE) samples as ground truth. There is usually a considerable shape mismatch between the freshly excised sample and its FFPE counterpart due to sample dehydration, and such mismatch negatively impacts the quality of the training data. We propose to address this challenge by using an innovative reliability-based training data selection method, where the reliability of the training data is quantified and estimated by using an unsupervised expectation maximization (EM) classification algorithm with soft probabilistic output. Experiment results demonstrate that the proposed multinomial Bayesian probit regression models with reliability-based training data selection achieve better performance than existing methods. Overall, these results demonstrate that the proposed supervised segmentation models represent a promising technique for the region detection with THz imaging of freshly excised breast cancer samples.

1. Introduction

Breast cancer is one of the most common forms of cancer in women across the U.S., with approximately 1 in 8 women diagnosed with breast cancer during their lifetime [1]. In 2021, the expected number of breast cancer cases is 281,550 with approximately 43,600 projected deaths in the U.S. alone [1]. Among feasible treatment options for early detected breast cancer, mastectomy and breast conserving surgery (BCS) are the customary care approaches. For instance, in BCS the cancerous tumor surrounded by a small margin of healthy breast tissue is removed. The evaluation of the margins in the excised sample is performed by a pathologist, who analyzes its formalin-fixed, paraffin-embedded (FFPE) representation. Since the histopathology process takes around 10–15 days, the re-excision rates of BCS oscillate between 20–30% [2]. Even though the pathology analysis of the sample is considered the gold standard in cancer detection, it is necessary to accelerate the margin assessment process of the mass such that it can be performed in the operating room without sacrificing the overall cancer detection accuracy. This necessitates the development of a computational-based imaging benchmark for the detection of breast cancer within freshly excised samples, such that the surgeon can evaluate the margins of freshly excised tissue in the operating room to reduce re-excision rates.

Terahertz (THz) imaging has shown great potential for material characterization in a vast variety of applications, such as integrated circuit inspection [3], security screening [4], food inspection [5], and biomedical applications [6–12]. The common objective across these studies is the classification of the reflected THz pulse into a fixed number of categories, but with different segmentation techniques based on unsupervised or supervised learning methods. In general, unsupervised learning algorithms, such as mixture models [13,6], and Fuzzy C-means [10], make inferences on patterns among the input observations without utilizing a training stage. These techniques are useful for initial data

* Corresponding author.
E-mail addresses: tachavez@uark.edu (T. Chavez), nvohra@uark.edu (N. Vohra), kbailey1@illinois.edu (K. Bailey), magda@uark.edu (M. El-Shenawee), wuj@uark.edu (J. Wu).

https://doi.org/10.1016/j.bspc.2021.102949
Received 10 December 2020; Received in revised form 24 June 2021; Accepted 28 June 2021

exploration, but could be limited by their model definition and the lack of prior information. On the other hand, supervised learning algorithms utilize a fraction of the ground truth information to capture intrinsic links among the predictors and responses, which can be exploited during the segmentation process. Some commonly used supervised segmentation techniques in medical imaging segmentation include support vector machine (SVM) [8,14,15], partial least squares-discriminant analysis (PLS-DA) [16,15], K-nearest neighbors [8,14,15], random forest [8,17], and convolutional neural networks (CNN) [3,4,18,5]. Although supervised learning algorithms have achieved favorable results in segmentation tasks for biomedical applications, the requirement of a large amount of training observations represents one of the main challenges for their implementations.

The requirement of large amount of training data is mainly due to high model complexity in most supervised learning methods. In THz imaging, each pixel corresponds to a high-dimensional THz pulse, which contains valuable information about the characterization of the material in its corresponding location. Direct processing of the high-dimensional THz pulse will result in a high model complexity. Hence, it is essential to identify the most relevant features embedded in the THz waveforms to achieve good segmentation performance while maintaining lower model complexity to reduce the amount of training data. To tackle this problem, the absorption coefficient and refractive index spectra per pixel are used by [9] as their most significant features for the region segmentation within human gastric tissues. As an alternative to pre-defined characteristics, it is possible to automatically identify the critical information-bearing features through dimension reduction approaches, such as principal component analysis (PCA) [19,20,15], and the low-dimensional ordered orthogonal projection (LOOP) [6,7] algorithm. Once the most relevant features are identified, the segmentation algorithm utilizes these attributes to perform inferences on the parameters of their discriminating models.

This paper introduces a novel supervised image segmentation algorithm for the detection of breast cancer in THz imaging of BCS samples. The proposed method is developed by using a multinomial Bayesian ordinal probit regression model with a reliability-based training data selection method. This proposed method differs from conventional probit regression algorithms with linear regression models [21,7] or binary classifications [22]. Two non-linear regression models, polynomial regression and kernel regression with random Fourier features (RFF) [23], are employed in the proposed method to establish the link between THz data and classification latent variables. Since the Bayesian regression algorithm relies on the estimation of a small number of model parameters, the size of the training set required for this task is considerably smaller than alternative machine learning approaches, such as CNN and random forest. This fact is particularly important for our analysis because the procurement of biomedical samples corresponds to a laborious process that involves clinical protocols, and multi-disciplinary collaborations. As a result, this type of research usually presents a limited number of specimens, which should be strategically employed to validate the study's findings. Hence, one of the main advantages of the proposed algorithm is the reduced number of training observations required for its model estimation, which is much less than deep learning approaches.

Unlike alternative studies that use FFPE homogeneous breast cancer samples [8,24], this paper employs freshly excised murine-derived heterogeneous samples, i.e. tumors that contain different regions, such as cancer, fibro, fat, etc. For training purposes, the ground truth information is collected from the histopathology analysis of the sample, which represents the gold standard of cancer detection and is obtained after the histopathology process of the tissue. Due to dehydration during the histopathology process, there is a significant shape mismatch between the fresh sample and its FFPE counterpart. The proposed method tackles this problem by utilizing a mesh morphing algorithm that reshapes the contour of the pathology results into the shape of the fresh sample [25]. To account for possible errors during the morphing process, we propose a new reliability-based training data selection method, which measures the reliability of training data by using the probabilistic output of an unsupervised expectation maximization (EM) method with Gaussian mixture models (GMM). Only data with reliability exceeding a certain threshold will be included in the training data set to ensure the quality of model training.

The rest of the article is organized as follows. Section 2 introduces the THz system and the procedure to collect the images. Section 3 presents the proposed regression model, and its training and testing procedures. Section 4 shows the experimental results. Section 5 concludes this study.

2. Materials and methods

This section describes the methodology to inject tumors in C57BL/6 black laboratory xenograft mice and the procedure to perform imaging using the THz system. The mice were kept on a high-fat diet until reaching a target weight of 35 g. At this point, the mice were injected with E0771 murine-derived breast adenocarcinoma cells to develop the tumors. Once the tumors reached a 1 cm diameter, they were excised under anesthesia [26]. The excised tumors were immersed in phosphate-buffered saline (PBS) solution to be transferred from the excision site to the THz lab for imaging using the THz system.

The TPS Spectra 3000 THz pulse reflection imaging system (TeraView, Ltd., UK) at the University of Arkansas was used [26]. The system uses a 780 nm Ti: Sapphire laser signal directed onto the THz antennas to generate the THz pulse. The samples handled in this work are measured in reflection mode, where the reflected signal was collected at every 200 μm size pixel on the tumor. This was achieved by placing the tumor onto the THz system scanner, which was set to increment at every 200 μm step size using stepper motors. The system was purged with dry nitrogen gas for 30 min prior to imaging to remove any water vapors in the core chamber.

The tumors to be imaged were prepared by drying any excessive fluid flowing out using filter paper, as shown in Fig. 1a. Then the tumor was placed between two polystyrene plates with a gentle pressure from the top to keep the imaging surface as flat as possible, as shown in Fig. 1b. This tumor arrangement is then placed on the scanning window for the imaging process, as shown in Fig. 1c [27]. After the imaging process, the tumors were immersed in formalin and sent to the Oklahoma Animal Disease Diagnostic Laboratory (OADDL) for the histopathology process.

All animals received care according to the Guide for the Care and Use of Laboratory Animals. In addition, the experimental process followed in this study was approved by the Institutional Animal Care and Use Committee (IACUC) of the University of Arkansas.

3. Theory and algorithm

3.1. Data pre-processing

This section describes the data pre-processing step, which is applied to the data prior to the training and testing procedures. The THz image can be represented by a third order tensor $\mathbf{V} \in \mathscr{R}^{N_1 \times N_2 \times F}$, with the first two dimensions representing the location of the pixel along the x and y axes with size N_1 and N_2, respectively, and the third dimension representing the frequency domain with size F. After unfolding, the THz information can be arranged in terms of a matrix $\mathbf{v} = [\mathbf{v}_1, \ldots c, \mathbf{v}_{N_s}]$, where $\mathbf{v}_n \in \mathscr{R}^F$ represents the amplitude of the frequency domain spectrum of the reflected waveform in the n-th pixel, and $n = \{1, \ldots c, N_s\}$ with $N_s = N_1 N_2$ corresponding to the total number of pixels in the THz image. The frequency domain response per pixel is a high-dimensional waveform of length $F = 106$ samples, which covers the system's operation range from 0.1 to 4 THz.

Before performing the image segmentation algorithm, we apply the LOOP algorithm [6] to the data to achieve dimension reduction.This

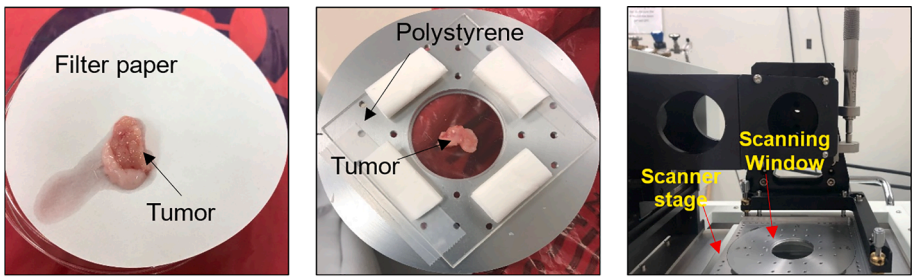

Fig. 1. Sample preparation for THz imaging. (a) Tumor placed on filter paper for drying excess fluid, (b) Tumor placed between two polystyrene plates, (c) Scanning window on the scanner stage upon which the tumor arrangement in (b) is positioned.

method projects the F-dimension signal per pixel into a lower-dimensional subspace of size $L < F$, which contains the most relevant features embedded in THz imaging waveforms.

The lower dimensional data at the output of the LOOP algorithm is then normalized, such that the features are scaled to zero mean and unit standard deviation. This procedure is repeated for all the samples in the data set. The normalized lower dimension data vector is represented by a row vector $\mathbf{x}_n \in \mathscr{R}^{1\times L}$, where $n = \{1, \ldots c, N\}$ and N corresponds to the total number of training observations. It is important to highlight that the training stage selects an equal number of observations per region to avoid bias in the trained model. Details about how the training samples are selected within the training data set are given in Section 3.3.

3.2. Multinomial Bayesian learning with probit regression

This section develops multinomial Bayesian ordinal probit regression models of the data, which are used to classify each pixel in the THz image to a certain region. Conventional probit regression models are commonly used in binary classification problems. We introduce a multi-class extension of this method that employs a continuous latent variable, $\mathbf{z} \in \mathscr{R}^N$, for non-binary partitions of the data set [21].

Given the estimated value of the latent variable, and a set of estimated thresholds, $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \ldots c, \alpha_K\}$, the region label per pixel is determined based on the range where the latent variable is located within $\boldsymbol{\alpha}$, e.g. the n-th pixel corresponds to the k-th region if $\alpha_{k-1} < z_n < \alpha_k$.

Two non-linear regression models are employed for the multinomial probit regression modeling of the data, and they are polynomial regression and kernel regression. We will introduce both models in this section, and compare the performance between the two different models in the section of experiment results.

3.2.1. Polynomial regression

In the polynomial regression model, the latent variables, $\{z_n\}_{n=1}^N$, are modeled as independent but non-identically distributed Gaussian random variables with variance σ^2. The mean of z_n is modeled as a Q-order polynomial regression of the L-dimensional data $\mathbf{x}_n$. The polynomial regression model can be represented as

$$z_n \overset{\text{ind}}{\sim} \mathscr{N}\left(\mathbf{w}_n\boldsymbol{\beta}, \sigma^2\right), \tag{1}$$

where $\mathbf{w}_n = [1, \mathbf{x}_n, \mathbf{x}_n^{(2)}, \ldots c, \mathbf{x}_n^{(Q)}] \in \mathscr{R}^{1\times(QL+1)}$, with $\mathbf{x}_n^{(k)}$ representing the element-wise k-th exponent of $\mathbf{x}_n$, $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots c, \beta_{QL}]^T$ is the regression parameter vector, and L is the dimension of the row vector $\mathbf{x}_n$. In this paper, we consider a fixed variance $\sigma^2 = 1$ in the polynomial regression model. The regression parameter $\boldsymbol{\beta}$ can be obtained through training, with details described in the next section.

3.2.2. Kernel regression

In the kernel regression model, the data vector of each pixel is mapped onto a higher, or even infinite, dimensional space as $h(\mathbf{x}_n)$, where $h : \mathbf{x}_n \in \mathscr{R}^L \rightarrow h(\mathbf{x}_n) \in \mathscr{R}^U$ represents the feature mapping, with $U > L$. With the kernel trick in the dual problem definition of the kernel regression model, it is not necessary to explicitly define the mapping function $h(\mathbf{x}_n)$ or the high-dimensional mapping space. Instead, the information per pixel is implicitly mapped by using a kernel function that represents the inner product between the two mapped vectors as

$$\mathscr{K}(\mathbf{x}_m, \mathbf{x}_n) = h(\mathbf{x}_m)h(\mathbf{x}_n)^T.$$

In this paper, the squared exponential kernel is used to model the inner product in the higher-dimensional mapping space as

$$\mathscr{K}(\mathbf{x}_m, \mathbf{x}_n) = e^{-\nu||\mathbf{x}_m - \mathbf{x}_n||^2} \tag{2}$$

where ν is the kernel parameter.

The complexity of the kernel regression model increases with the size of the training data set. The number of training samples used in this study is in general much smaller than other supervised learning algorithms such as deep learning. However, there is still a large number of pixels within each case that can negatively impact the model complexity. We propose to further reduce model complexity by using a random Fourier features (RFF) approximation [23], which can reduce the number of parameters that need to be estimated during the training process. The RFF method explicitly projects the vectors per pixel into a lower dimensional approximation of the kernel's feature space as $h_{\text{RFF}}(\mathbf{x}_n)$, where $h_{\text{RFF}} : \mathbf{x}_n \in \mathscr{R}^L \rightarrow h(\mathbf{x}_n) \in \mathscr{R}^V$ with $V < U$ and

$$\mathscr{K}(\mathbf{x}_m, \mathbf{x}_n) \approx h_{\text{RFF}}(\mathbf{x}_m)h_{\text{RFF}}(\mathbf{x}_n)^T. \tag{3}$$

In order to obtain h_{RFF}, we can express the shift-invariant kernel functions by following Bochner's theorem as

$$\mathscr{K}(\mathbf{x}_m - \mathbf{x}_n) = \int_{\mathscr{R}^L} e^{i\boldsymbol{\omega}^T(\mathbf{x}_m - \mathbf{x}_n)} P(\boldsymbol{\omega}) d\boldsymbol{\omega} \tag{4}$$

where $P(\boldsymbol{\omega})$ corresponds to the Fourier transform of the kernel, and $\boldsymbol{\omega} \in \mathscr{R}^{L\times 1}$ is the vector corresponding to the frequency domain variable. Since it is not possible to directly compute (4), we employ a Monte Carlo approach by assuming that $P(\boldsymbol{\omega})$ takes the form of a probability distribution, with $\boldsymbol{\omega}$ following a multivariate Gaussian distribution of the form $P(\boldsymbol{\omega}) = \mathscr{N}(\mathbf{0}_L, 2\nu\mathbf{I}_L)$. By following the Monte Carlo approach, the kernel function in (4) can be approximated by

$$\mathscr{K}(\mathbf{x}_m - \mathbf{x}_n) \approx \frac{1}{Q}\sum_{q=1}^{Q}\begin{pmatrix}\cos(\boldsymbol{\omega}_q^T\mathbf{x}_m)\\ \sin(\boldsymbol{\omega}_q^T\mathbf{x}_m)\end{pmatrix}^T\begin{pmatrix}\cos(\boldsymbol{\omega}_q^T\mathbf{x}_n)\\ \sin(\boldsymbol{\omega}_q^T\mathbf{x}_n)\end{pmatrix},$$

where $\boldsymbol{\omega}_q \overset{\text{iid}}{\sim} P(\boldsymbol{\omega})$, and Q is the total number of Monte Carlo iterations [23]. Through this expression, the feature space defined by RFF can then be expressed as

$$h_{\mathrm{RFF}}(\mathbf{x}) = \frac{1}{\sqrt{Q}}\begin{bmatrix}\cos(\boldsymbol{\Omega}^T\mathbf{x})\\ \sin(\boldsymbol{\Omega}^T\mathbf{x})\end{bmatrix} \in \mathscr{R}^{2Q\times 1} \tag{5}$$

where $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_Q] \in \mathscr{R}^{L\times Q}$. In (5), the L-dimension data vector $\mathbf{x}$ is projected onto a feature space of dimension $V = 2Q$. The number of Monte Carlo iterations can be set according to a fixed error per entry, $\pm\zeta$, where $Q = \log(N)/\zeta^2$, or in general as, $Q = \sqrt{N}\log(N)$ [23].

The latent variable for the n-th pixel can be modeled as

$$z_n \sim \mathscr{N}\left(\mathbf{w}_n\boldsymbol{\beta}, \sigma^2\right), \tag{6}$$

where $\mathbf{w}_n = h_{\mathrm{RFF}}(\mathbf{x}_n)^T \in \mathscr{R}^{1\times 2Q}, \sigma^2 = 1$, and the vector $\boldsymbol{\beta} \in \mathscr{R}^{2Q\times 1}$ contains the regression coefficients to be estimated through the training process.

3.3. Training process

This section describes the newly proposed reliability-based training data selection method, and the training process of the model parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, with a Markov chain Monte Carlo (MCMC) method.

3.3.1. Reliability-based training data selection

The training step utilizes 6 murine fresh samples with the same number of regions, including cancer, fibro or muscle, and fat. The regions in the THz images are labeled by using pathology results. Since the fresh tissue goes through a dehydration process during the pathological analysis, there is a considerable mismatch between the region allocations of fresh tissues and the corresponding pathology image. To correct this mismatch, we utilize a mesh morphing algorithm to reshape the contour of the pathology results into the shape of the THz image taken from the freshly excised sample [25]. The mesh morphing algorithm matches the pathology and THz images by using control points on the contour of the tissue, thus it is possible that there is still internal mismatch between the two images after morphing. As a result, some of the pixels in the training THz images might be erroneously labeled due to the residual mismatch with the pathology image. Therefore, it is important to quantify the reliability of the ground truth information to avoid the usage of erroneously labeled pixels as training observations.

We propose to measure the reliability of the ground truth information for each pixel by using the results obtained through an unsupervised Bayesian learning approach with GMM and EM [6]. The output of the unsupervised EM algorithm contains the probability that each pixel belongs to a certain region. A pixel will be selected for the training data set only if the probability exceeds a certain threshold, and the corresponding region matches the pathology results. In this article, the probability threshold selected for this procedure was 60%. Thus the unsupervised results serve as a reliability indicator for the morphed pathology image, which reduces error in the training procedure.

3.3.2. Parameter initialization

Before starting the iterative MCMC training process, we need to obtain the initial values of the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

To ensure that the $\boldsymbol{\alpha}$ parameter covers the entire latent variable domain, $\mathscr{R}$, certain elements within this parameter are manually fixed as $\alpha_0 = -\infty, \alpha_1 = 0$, and, $\alpha_K = \infty$ [21]. Thus the probability that the n-th pixel belongs to the first region is as follows,

$$\Pr(y_n = 1) = \Phi(\alpha_1 - \mathbf{w}_n\boldsymbol{\beta}) - \Phi(\alpha_0 - \mathbf{w}_n\boldsymbol{\beta}) = \Phi(-\mathbf{w}_n\boldsymbol{\beta}),$$

or equivalently

$$-\mathbf{w}_n\boldsymbol{\beta} = \Phi^{-1}[P(y_n = 1)],$$

where Φ^{-1} corresponds to the inverse of the cumulative standard Gaussian distribution, and $\Pr(y_n = 1)$ is from the pathology results. It is possible to further rewrite this expression by utilizing its vector representation,

$$\mathbf{q} = -\mathbf{W}\boldsymbol{\beta}, \tag{7}$$

where $\mathbf{W} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \ldots, \mathbf{w}_N^T]^T, \mathbf{q} = [q_1,\ldots,q_N]^T \in \mathscr{R}^{N\times 1}$ with $q_n = \Phi^{-1}(\Pr(y_n = 1))$.

The parameter $\boldsymbol{\beta}$ can then be initialized by using the least squares (LS) estimate as

$$\boldsymbol{\beta} = -(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{q}. \tag{8}$$

In the ground truth data obtained from the pathology results, $\Pr(y_n = 1)$ can take two values, 0 or 1, based on the pathology label. However, we cannot directly use these exact results in (8) because $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$. To address this problem, we assign $\Pr(y_n = 1) = 1 - \epsilon$ if the n-th pixel belongs to the first class in the pathology results, and $\Pr(y_n = 1) = \epsilon$ otherwise, with ϵ being a small number. In this paper we choose $\epsilon = 0.0013$.

Similar to the initialization process of the $\boldsymbol{\beta}$ parameter, we utilize the fixed elements within the $\boldsymbol{\alpha}$ parameter to estimate the remaining unknown elements within this vector, $\{\alpha_2, \ldots c, \alpha_{K-1}\}$. For this purpose, consider the following expression:

$$\begin{aligned}\Pr(y_n = K) &= \Phi(\alpha_K - \mathbf{w}_n\boldsymbol{\beta}) - \Phi(\alpha_{K-1} - \mathbf{w}_n\boldsymbol{\beta})\\ &= 1 - \Phi(\alpha_{K-1} - \mathbf{w}_n\boldsymbol{\beta}).\end{aligned}$$

Thus

$$\alpha_{K-1} = \mathbf{w}_n\boldsymbol{\beta} + \Phi^{-1}[1 - \Pr(y_n = K)].$$

The value of α_{K-1} can then be estimated by using the N training observations as,

$$\alpha_{K-1} = \frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{w}_n\boldsymbol{\beta} + \Phi^{-1}[1 - \Pr(y_n = K)]\right\}. \tag{9}$$

Since this paper explores the implementation of the probit regression approach for the segmentation of THz images with $K = 3$ regions, it was only necessary to find the element α_2 within these models. Alternatively, if $K > 3$, this process can be repeated to estimate the remaining unknown elements within the $\boldsymbol{\alpha}$ parameter by utilizing α_{K-1}.

3.3.3. Training with MCMC

Once the training set is selected and the parameters are initialized, we proceed to estimate the regression parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, through an MCMC process. The prior distributions of the model parameters $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ are defined as:

$$\begin{aligned}\pi(\boldsymbol{\alpha}) &= \prod_{k=1}^{K} 1(\alpha_k > \alpha_{k-1}),\\ \pi(\boldsymbol{\beta}) &= \mathscr{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0),\end{aligned}$$

with $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$ representing the hyper-parameters of this approach. In this paper, we consider $\boldsymbol{\beta}_0 = \mathbf{0}$, and $\boldsymbol{\Sigma}_0 = 10^4 \times \mathbf{I}$.

The estimation stage utilizes an MCMC process with the following posterior distributions [21]:

- Posterior distribution of $\mathbf{z}$,

$$z_n|\boldsymbol{\beta}, \boldsymbol{\alpha}, y_n = k \sim \begin{cases} 0 & z_n \leqslant \alpha_{k-1} \\ \dfrac{\phi(\mathbf{w}_n\boldsymbol{\beta}, \sigma^2; z_n)}{\Phi\left(\dfrac{\alpha_k - \mathbf{w}_n\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\dfrac{\alpha_{k-1} - \mathbf{w}_n\boldsymbol{\beta}}{\sigma}\right)} & \alpha_{k-1} < z_n < \alpha_k \\ 0 & z_n \geqslant \alpha_k \end{cases} \tag{10}$$

where $\phi(\mu, \sigma^2; x)$ represents the Gaussian probability density function (pdf) with mean μ and variance σ^2 evaluated in x; and $\Phi(x)$ is the cumulative distribution function (CDF) of a standard Gaussian

variable with 0 mean and unit variance.

- Posterior distribution of $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}|\mathbf{z} \sim \mathcal{N}(\mu_\beta, \boldsymbol{\Sigma}_\beta), \tag{11}$$

where $\boldsymbol{\Sigma}_\beta = \left[\left(\mathbf{W}^T\mathbf{W} + \sigma^2\boldsymbol{\Sigma}_0^{-1}\right)/\sigma^2\right]^{-1}$, and $\mu_\beta = \boldsymbol{\Sigma}_\beta\left[\left(\mathbf{W}^T\mathbf{z} + \sigma^2\boldsymbol{\Sigma}_0^{-1}\beta_0\right)/\sigma^2\right]$.

- Posterior distribution of $\boldsymbol{\alpha}$,

$$\alpha_k|\mathbf{z}, \mathbf{y}, \alpha_{j\neq k} \sim \mathcal{U}(a, b), \tag{12}$$

where $\mathcal{U}$ represents a uniform distribution with parameters $a = \max\left(\max\{z_n : y_n = k\}, \alpha_{k-1}\right)$, and $b = \min\left(\min\{z_n : y_n = k+1\}, \alpha_{k+l}\right)$.

Overall, the training procedure is summarized in Algorithm 1, where mod represents the modulo operator and M corresponds to the total number of MCMC iterations to be considered in the testing process. It is important to mention that the MCMC algorithm runs for a total of $10M$ iterations, where the first half are discarded during the burn-in period, and the regression parameters are stored every 5 iterations after this period. This operation leaves a total of M samples from the posterior distributions of the regression parameters, which are used during the testing procedure. In this paper, the results were produced by considering that $M = 4,000$, which results in a total of $40,000$ MCMC iterations.

Algorithm 1: Training procedure.

Input: Data $\mathbf{W}$, labels $\mathbf{y}$, hyperparameters $\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, \sigma^2$
Initialization: Estimate $\boldsymbol{\beta}$ and the unknown elements within $\boldsymbol{\alpha}$ using (8) and (9), respectively
for $j = 1, \ldots c, 10M$ **do**
  Draw $\mathbf{Z}^{(j)}$ from (10) using $\boldsymbol{\beta}^{(j-1)}, \boldsymbol{\alpha}^{(j-1)}$, and $\mathbf{y}$.
  Draw $\boldsymbol{\beta}^{(j)}$ from (11) using $\mathbf{Z}^{(j)}$.
  Draw the unknown elements within $\boldsymbol{\alpha}^{(j)}$ from (12) using $\mathbf{Z}^{(j)}$, and $\mathbf{y}$.
  if $j > 5M$ and $j \bmod 5 = 0$ **then**
    Store $\boldsymbol{\beta}^{(j)}$ and $\boldsymbol{\alpha}^{(j)}$.
  end if
end for
Output: Regression parameters $[\boldsymbol{\beta}^{(i)}, \boldsymbol{\alpha}^{(i)}]_{i=1}^{M}$.

3.4. Testing process

This section presents the testing procedure of the proposed multinomial probit regression algorithm. The algorithm is tested by using the THz images from samples not used during the training process. Similar to the training data, the data used for testing goes under the same preprocessing procedures, which include obtaining the frequency response of the pulse per pixel and dimension reduction.

Once the corresponding model parameters are obtained during the training phase, as described in Section 3.3, the region assignment is performed by using the following soft clustering scheme. Denote the parameters obtained through training in the i-th MCMC iteration as $\{\alpha_k^{(i)}\}_{k=0}^{K}$ and $\boldsymbol{\beta}^{(i)}$. With the multi-class probit regression algorithm, the latent variable of the n-th pixel in the testing data can be modeled by applying the model parameters from the i-th iteration of the MCMC training as

$$z_n^{(i)} \sim \mathcal{N}(\mathbf{w}_n\boldsymbol{\beta}^{(i)}, \sigma^2), \quad \text{for } i = 1, \ldots, M \tag{13}$$

Thus

$$\Pr(\alpha_{k-1}^{(i)} < z_n^{(i)} \leqslant \alpha_{k-1}^{(i)}) = \Phi\left(\frac{\alpha_k^{(i)} - \mathbf{w}_n\boldsymbol{\beta}^{(i)}}{\sigma}\right) - \Phi\left(\frac{\alpha_{k-1}^{(i)} - \mathbf{w}_n\boldsymbol{\beta}^{(i)}}{\sigma}\right) \tag{14}$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function.

The probability that the n-th pixel belongs to the k-th category can then be calculated as

$$\Pr(y_n = k) = \frac{1}{M}\sum_{i=1}^{M}[\Pr(\alpha_{k-1}^{(i)} < z_n^{(i)} \leqslant \alpha_{k-1}^{(i)})] \tag{15}$$

where M is the total number of stored MCMC iterations. With (15), we evaluate the likelihood of each pixel from the testing data with respect to every region in the tissue.

4. Experimental results

The experimental results are obtained by applying the proposed multinomial probit regression algorithm on the testing data. The training and testing data are obtained from freshly excised xenograft murine samples with 3 regions each, such as cancer, muscle or fibro, and fat. The samples correspond to mice 6B, 8B, 9A, 9B, 10A, 10B, and 13A. Samples 9B, 10B, and 13A are used for testing, and all remaining samples (6B, 8B, 9A, and 10A) are used for training exclusively. It is important to mention that each testing sample is also utilized for the training procedure of a different testing sample. For example, when testing sample 9B, we utilized the 6 remaining samples, 10B, 13A, 6B, 8B, 9A, and 10A to train its model. While the training and testing processes employ 6 and 3 samples, respectively, the overall amount of training pixels is smaller than its testing counterpart. As an example, mouse 9B utilized 3,192 pixels for training, and 4,797 pixels for testing. The number of training pixels is reduced due to the application of the reliability-based training selection process. Although some pixels are discarded, this step is crucial to avoid the utilization of mistakenly assigned ground truth pixels. In addition, some regions, such as muscle and fibro, are commonly smaller than the rest of the regions in a murine tumor sample. To avoid the introduction of bias in our model, the algorithm selects the same amount of pixels per region, which further reduces the total amount of training observations.

The results obtained from the proposed algorithms are compared with two previously published unsupervised learning approaches based on GMM, which are 1-dimensional (1D) MCMC [26] and 2-dimensional (2D) EM [6]. Source codes for the multinomial probit regression algorithm can be found in [28]. The quantitative analysis of the segmentation model is summarized through ROC curves, which identify the true vs. false positive detection rates per region. Since the proposed algorithms utilize a soft-clustering segmentation approach, the ROC curves represent the potential detection results that can be obtained by the selection of a suitable classification threshold. Details on the generation of the ROC curves can be found in Appendix A.

4.1. Data exploration

We first implemented a univariate t-Test to verify that there is a significant difference between the mean of cancerous vs. the mean of non-cancerous pixels within each testing sample. The test is performed by using the first component of the low-dimension vector per pixel at the output of the LOOP algorithm. The null hypothesis of the test is that the LOOP outputs of cancerous and non-cancerous pixels will have the same

Table 1
Paired-sample t-Test results with 0.05 significance level.

Sample	Test statistic	Degrees of freedom	Standard deviation	Confidence interval	p-value
Mouse 9B Fresh	19.0007	198	0.6702	[1.6139, 1.9877]	1.6625×10^{-46}
Mouse 10B Fresh	18.4410	198	0.7535	[1.7550, 2.1753]	7.3819×10^{-45}
Mouse 13A Fresh	31.4908	198	0.4205	[1.7555, 1.9900]	4.9522×10^{-79}

mean. The results of the t-Test are summarized in Table 1, where we can observe that the *p*-value is close to zero for all the testing samples. Such results reject the null hypothesis, therefore it is demonstrated through the t-Test results that there are significant differences between the mean of the LOOP outputs of cancerous and that of non-cancerous pixels.

Given the promising results of the t-Test and considering that the data is a 2 or 3 dimensional vector, we performed additional in-depth analysis by implementing a Hotelling T-squared test. Unlike the univariate case, this test utilized the full vector per pixel at the output of the LOOP algorithm. The test hypothesis is the same as the t-Test. The results of this technique are summarized in Table 2, where we can observe that the *p*-value is close to zero for all the testing samples. Such results reject the null hypothesis for the multivariate case.

To further illustrate the results of these tests, we have plotted the empirical marginal probability density function (PDF) of one sample, Mouse 9B fresh, as shown in Fig. 2. Features 1 and 2 correspond to the first 2 components of the low-dimension vector per pixel obtained through the dimension reduction algorithm, LOOP. It is important to clarify that these features do not correspond to a specific physical feature within the THz waveform, instead they represent a combination of intrinsic key characteristics within the waveform that are automatically found by the dimension reduction technique. From this plot, we can observe that the distribution of the cancerous region (the red plot) is different from those of the non-cancerous regions (green and blue plots) in at least one dimension. In particular, the fat region is significantly different than the cancer region, while muscle presents some minimal vicinity to the cancer mean. These plots also verify that the overall PDF of the data resembles a Gaussian distribution.

4.2. Mouse 9B fresh

The first sample is mouse 9B fresh, which contains 3 regions: cancer, muscle, and fat. The THz image of this sample is shown in Fig. 3a, which was procured while the tissue was still fresh. This figure utilizes the power spectra of the reflected THz waveform as the summarization feature per pixel. It can be observed here that the cancer region (red color) in the sample shows higher reflection than the surrounding fat tissue (blue color). However, the differentiation between the muscle and cancer regions is not so obvious. This could be because the electrical properties of muscle and cancer are identical in the THz range [26]. Fig. 3b represents the pathology analysis of this sample, which clearly indicates the location and the extent of the regions within the tissue. Fig. 3c shows the morphed pathology results obtained from the mesh morphing algorithm [25]. Figs. 3d and e correspond to the 1D MCMC [26] and 2D EM [6] segmentation results, respectively. Finally, Figs. 3f and g represent the multinomial probit segmentation results obtained by using the 3D polynomial and kernel regression models, respectively. It is important to mention that these models' results were obtained by utilizing the optimal segmentation thresholds of each ROC curve, which prioritized the detection of cancer among all regions followed by muscle or fibro. For the supervised regression models, the algorithm utilizes 6 murine fresh samples within its training information, which correspond to mice 6B, 8B, 9A, 10A, 10B, and, 13A. In addition, the polynomial regression model employs a fifth order polynomial definition, and the kernel regression model uses $\nu = 0.3$ and RFFs with $Q = 20$.

Table 2
Hotelling T-squared test with 0.05 significance level.

Sample	Hotelling's T-Squared statistic	Degrees of freedom	Approximation statistic test (χ^2)	*p*-value
Mouse 9B Fresh	661.6546	3	661.6546	0.0000
Mouse 10B Fresh	401.8756	3	401.8756	0.0000
Mouse 13A Fresh	1189.6145	2	1189.6145	0.0000

By visually inspecting the images, we can observe that there is a good correlation between the detection results and the morphed pathology results regarding the regions of cancer and fat. There is misclassification in the muscle area for all three algorithms, and the 1D MCMC model presents the largest misclassification of this region.

To quantitatively evaluate these results, we introduce the ROC curves of all the segmentation models in Fig. 4. The ROC curves show the true detection rate as a function of false detection rate. Regarding cancer and fat, all multivariate detection approaches, that is, 2D EM (unsupervised), 3D polynomial regression (supervised), and 3D kernel regression (supervised), achieve similar performance, regardless whether they are supervised or unsupervised approaches. The performance of the 1D MCMC algorithm is worse than its multivariate counterparts for both the cancer and fat regions. The advantage of the supervised approach is demonstrated in the ROC curve for the muscle region, where it is observed that the two proposed probit algorithms (3D polynomial regression and 3D kernel regression) achieve significant performance gain over the two unsupervised algorithms.

This performance gain can be quantified by analyzing the areas under the ROC curves, which are shown in Table 3. An ideal classifier with 0 false detection rate and 100% sensitivity (true detection rate) achieves a 100% area under its ROC curve. In this table, we can observe that the supervised regression models proposed in this paper obtain the largest areas under the ROC curves for all regions, with muscle representing the highest performance gain from 71.35% to 86.80%.

4.3. Mouse 13A fresh

The second sample is mouse 13A fresh, which contains 4 regions: cancer, fibro, fat, and a lymph node. Since the lymph node in this sample shows signs of metastasis, we consider its area as part of the cancer region in the morphed pathology image. Therefore, the total number of regions considered for the segmentation task of this sample is 3: cancer, fibro, and fat. Fig. 5a represents the THz image that was collected while the tissue was fresh. Similar to the previous sample, we observe that cancer (red color) shows higher reflection than fat (blue color). Figs. 5b and c correspond to the histopathology analysis of the tissue and its corresponding morphed mask, respectively. Figs. 5d and e represent the results obtained through the unsupervised Gaussian mixture models. The linear and kernel regression models are represented in Figs. 5f and g, respectively. For the analysis of this sample, the supervised learning techniques utilize 6 murine fresh samples for its training step, which correspond to: 6B, 8B, 9A, 9B, 10A, and 10B. Furthermore, the polynomial regression utilizes a first order polynomial representation, and the kernel regression model uses $\nu = 0.1$ and RFFs with $Q = 20$.

The ROC curves of the classifiers are shown in Fig. 6, where we can observe that the cancer and muscle detection performance improves by using the 2D supervised linear regression model. This can be further confirmed in Table 3, where we can observe that the area under the cancer ROC curve improves from 86.38% to 93.23% by using the supervised linear regression algorithm. Similarly, the area under the fibro ROC curve increases from 72.63% to 78.10%.

4.4. Mouse 10B fresh

Finally, the third sample is mouse 10B fresh, which contains 3 regions: cancer, muscle, and fat. Fig. 7a represents the THz image of this sample. Figs. 7b and c correspond to the pathology analysis and its morphed representation, respectively. A wide gap between the cancer region as seen in the pathology image is due to the lumens in the cancer. When fresh, these lumens were filled with fluid secretions. Hence, it can be observed that the lumens in cancer show higher reflection than the rest of the region, which are presented in dark red within Fig. 7a. Figs. 7d and e represent the unsupervised classification results obtained

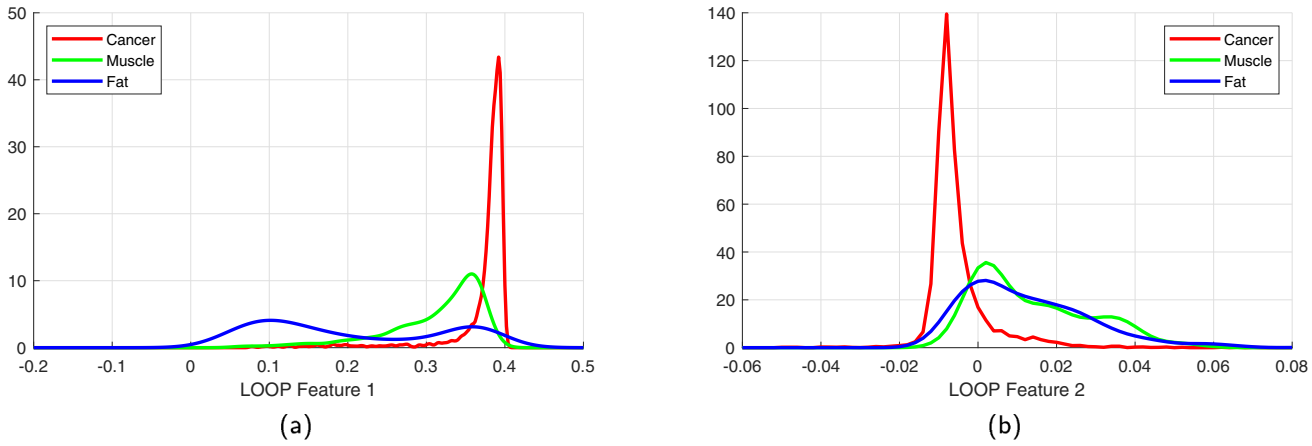

Fig. 2. Estimated marginal probability density function (PDF) of Mouse 9B fresh (a) LOOP Feature 1. (b) LOOP Feature 2.

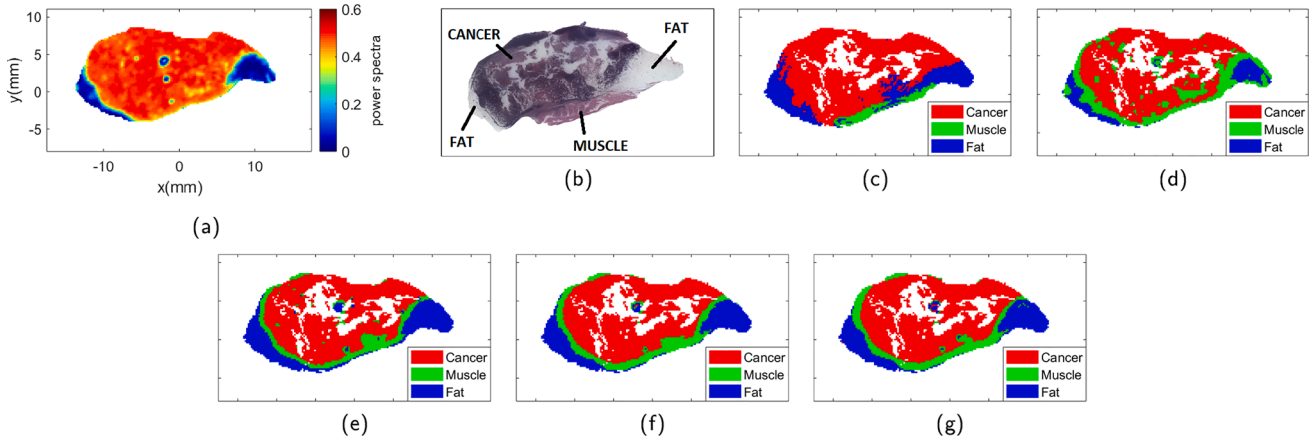

Fig. 3. Sample Mouse 9B Fresh. (a) THz image [29]. (b) Pathology image [29]. (c) Morphed Pathology [29]. (d) 1D MCMC model [29]. (e) 2D unsupervised EM model. (f) 3D supervised polynomial regression model (this work). (g) 3D supervised RFF kernel mo.del (this work).

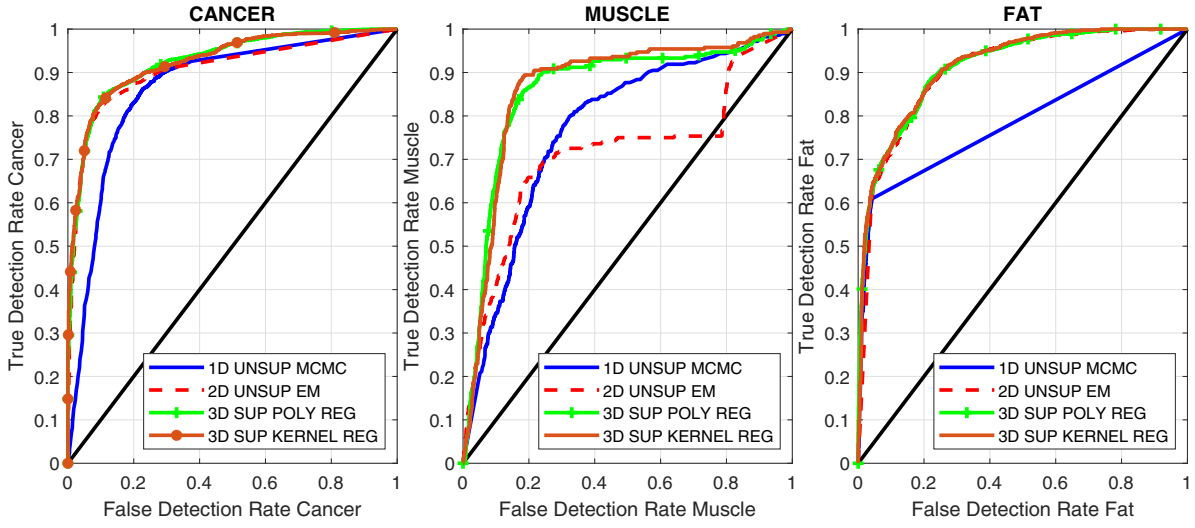

Fig. 4. ROC curves for sample Mouse 9B Fresh.

through the 1D MCMC and 2D EM approaches, respectively. Figs. 7f and g illustrate the segmentation results obtained through the supervised linear and kernel regression models, respectively. For the supervised regression models, the algorithm utilizes 6 murine fresh samples for its training step, which correspond to mice 6B, 8B, 9A, 9B, 10A, and, 13A. Additionally, the polynomial regression approach employs a first order polynomial definition, and the kernel regression model uses $\nu = 0.64$ and RFFs with $Q = N\log(N) = 442$.

Table 3
Areas under the ROC curves.

Mouse 9B Fresh				
Region	1D MCMC	2D unsupervised EM	3D supervised polynomial regression	3D supervised kernel regression
Cancer	0.8647	0.9068	0.9271	0.9263
Muscle	0.7707	0.7135	0.8618	0.8680
Fat	0.7874	0.9066	0.9144	0.9158
Mouse 13A Fresh				
Region	1D MCMC	2D unsupervised EM	2D supervised linear regression	2D supervised kernel regression
Cancer	0.8587	0.8638	0.9323	0.8909
Fibro	0.6637	0.7263	0.7810	0.7503
Fat	0.8626	0.9159	0.9288	0.8840
Mouse 10B Fresh				
Region	1D MCMC	2D unsupervised EM	2D supervised linear regression	3D supervised kernel regression
Cancer	0.7340	0.7894	0.8167	0.7732
Fibro	0.5539	0.6970	0.7525	0.7000
Fat	0.8970	0.9363	0.9468	0.9096

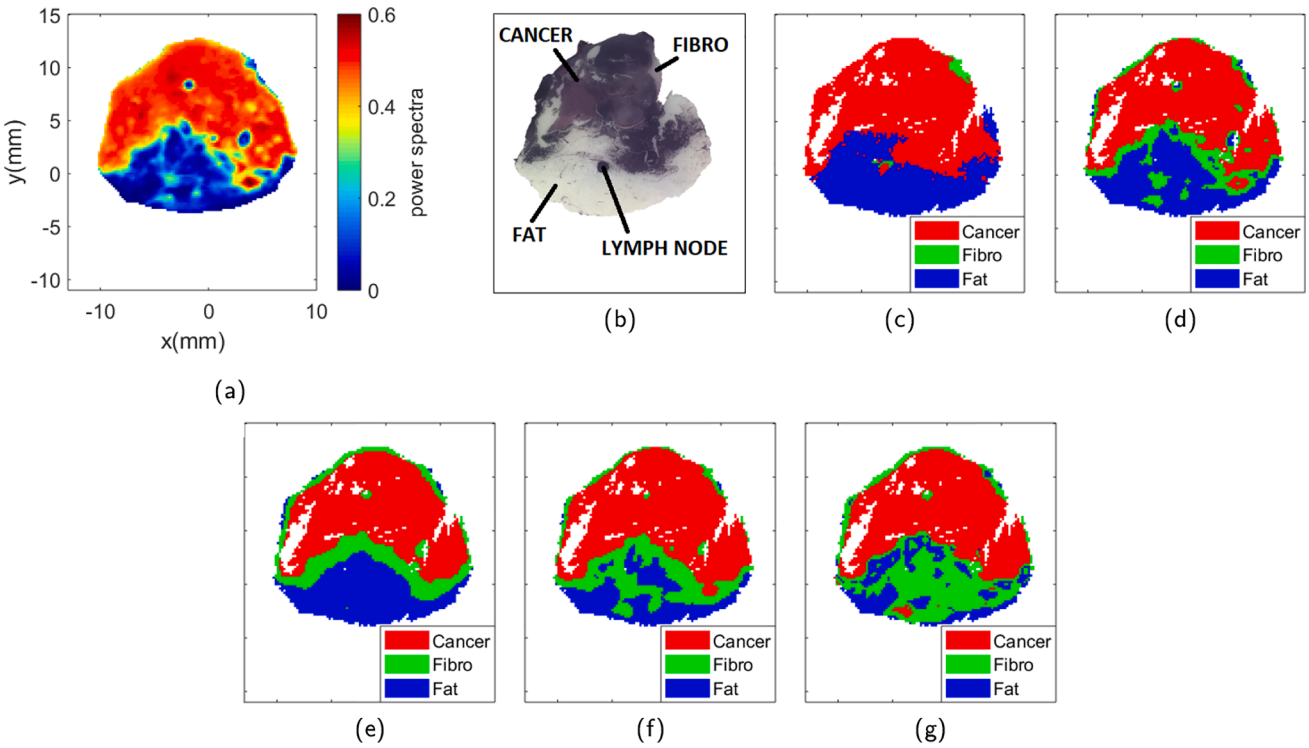

Fig. 5. Sample Mouse 13A Fresh. (a) THz image [25]. (b) Pathology image [25]. (c) Morphed Pathology [25]. (d) 1D MCMC model [25]. (e) 2D unsupervised EM model. (f) 2D supervised linear regression model (this work). (g) 2D supervised RFF kernel mo.del (this work).

The quantitative evaluation of the results are shown in Fig. 8 in the form of ROC curves. Similar to the previous samples, the ROC curves of the supervised models achieve better classification results. In particular, the 2D supervised linear regression model presents the best overall classification results among the tested classifiers. This can be further confirmed in Table 3, where we can observe that the areas under the cancer and muscle ROC curves increases from 78.94% to 81.67%, and 69.70% to 75.25%, respectively, when employing the proposed supervised segmentation model.

4.5. Comparison to SVM

To verify that the proposed algorithm significantly reduces the computational complexity of the training procedure, we compare the results of the proposed classifiers with respect to SVM. For fairness of comparison, we do not implement any dimension reduction or reliability-based training selection processes for the SVM classifier. As shown in Table 4, the computational time for the training procedure of the proposed classifier is lower than SVM, with SVM taking 30–36 min and the probit regression approach taking 1 min for most cases. It is important to clarify that the kernel regression implemented for Mouse 10B takes approximately 37 min due to the large amount of parameters that were estimated, where $Q = 442$. Hence, the proposed classifier can potentially reduce the training time as long as the number of parameters is set to a smaller amount, as is the case for $Q = 20$.

The segmentation results of the SVM model are further compared to the proposed kernel regression classifier in Fig. 9. In particular, we can observe that while the SVM approach can potentially detect the cancer and fat regions, it fails to detect the muscle region completely in Fig. 9b. Additionally, the quantitative segmentation results of the SVM classifier

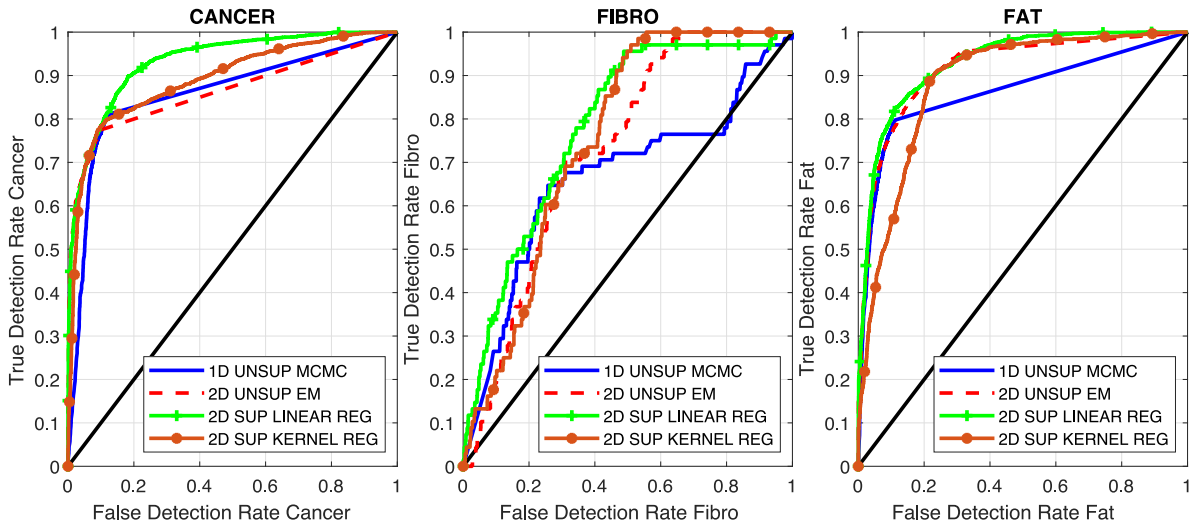

Fig. 6. ROC curves for sample Mouse 13A Fresh.

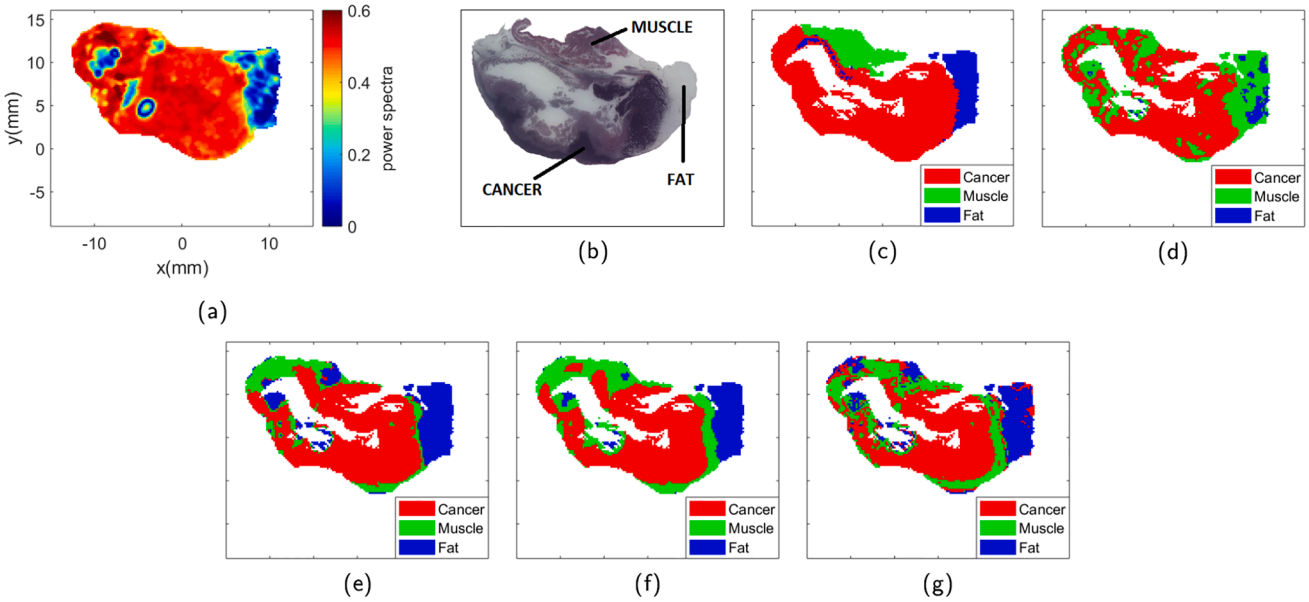

Fig. 7. Sample Mouse 10B Fresh. (a) THz image. (b) Pathology image. (c) Morphed Pathology. (d) 1D MCMC model. (e) 2D unsupervised EM model. (f) 2D supervised linear regression model (this work). (g) 3D supervised RFF kernel model (this work).

are summarized in Fig. 10. Since an SVM classifier is a hard-clustering technique, the performance of this classifier is represented as single points within the ROC curves. These results further confirm that the proposed classifiers present better segmentation results than a well-known technique such as SVM.

5. Conclusions

We have proposed a supervised multinomial Bayesian learning method for cancer detection using THz imaging of freshly excised samples. This algorithm utilizes multinomial Bayesian ordinal probit regression models to perform region classifications in THz images. Two probit regression models, a polynomial regression model and a kernel regression model, are adopted to represent the link between the THz features and their corresponding classification results. The proposed supervised learning approach requires considerably less amount of training data than other supervised learning approaches, such as CNN. During the training phase, in order to account for the mismatch between THz image and pathology results caused by deformation of the tissue during its histopathology process, we have proposed a reliability-based training data selection method, and only data that exceed a certain reliability threshold are used for training. Experimental results demonstrated that the proposed supervised regression models outperform existing algorithms, such as 1D MCMC and 2D EM, for all regions of interests. For instance, the areas under the cancer and muscle ROC curves in Mouse 9B fresh increases from 90.68% to 92.71%, and 71.35% to 86.18%, respectively, when utilizing the supervised polynomial regression approach.

In general, the supervised polynomial regression model obtained the highest areas under the ROC curves among all the presented classifiers, followed by the kernel regression model. In terms of the muscle and fibro region, we can highlight that the proposed supervised segmentation models achieve a considerable area increase when compared with their unsupervised counterparts, from 69.70% −72.63% to 75.25% −86.18%. These results represent a step forward towards the optimal differentiation between cancer vs. non-cancerous tissue within freshly excised BCS

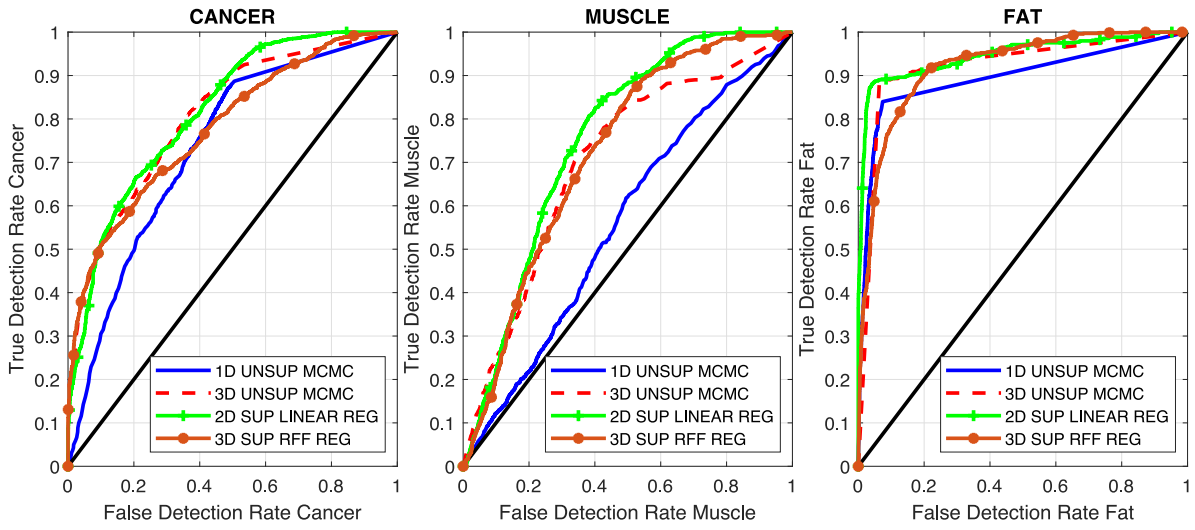

Fig. 8. ROC curves for sample Mouse 10B Fresh.

Table 4
Comparison of computational time for the training process.

Sample	SVM	Polynomial regression	Kernel regression
Mouse 9B	30.5955 min.	1.1254 min.	1.2918 min.
Mouse 13A	31.8640 min.	0.8458 min.	0.7604 min.
Mouse 10B	35.9760 min.	0.7333 min.	36.8386 min.

samples. In the mean time, it is recognized that achieving the areas under ROC curves to at least 90% for all regions still remains a challenge, and we plan to further improve the performance by developing higher dimensional latent variables in our future work.

CRediT authorship contribution statement

Tanny Chavez: Methodology, Software, Validation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Nagma Vohra:** Methodology, Investigation, Resources, Data curation, Writing - review & editing. **Keith Bailey:** Methodology, Investigation, Resources, Data curation, Writing - review & editing. **Magda El-Shenawee:** Conceptualization, Methodology, Resources, Supervision, Writing - review & editing, Funding acquisition. **Jingxian Wu:** Conceptualization, Methodology, Investigation, Resources, Supervision, Writing - original draft, Writing - review & editing, Funding acquisition.

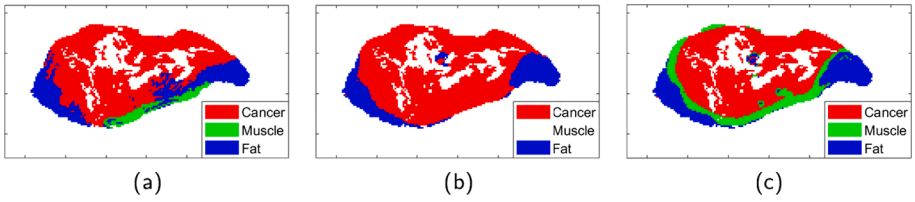

Fig. 9. Sample Mouse 9B Fresh. (a) Morphed Pathology [29]. (b) SVM model. (c) 3D supervised RFF kernel model (this work).

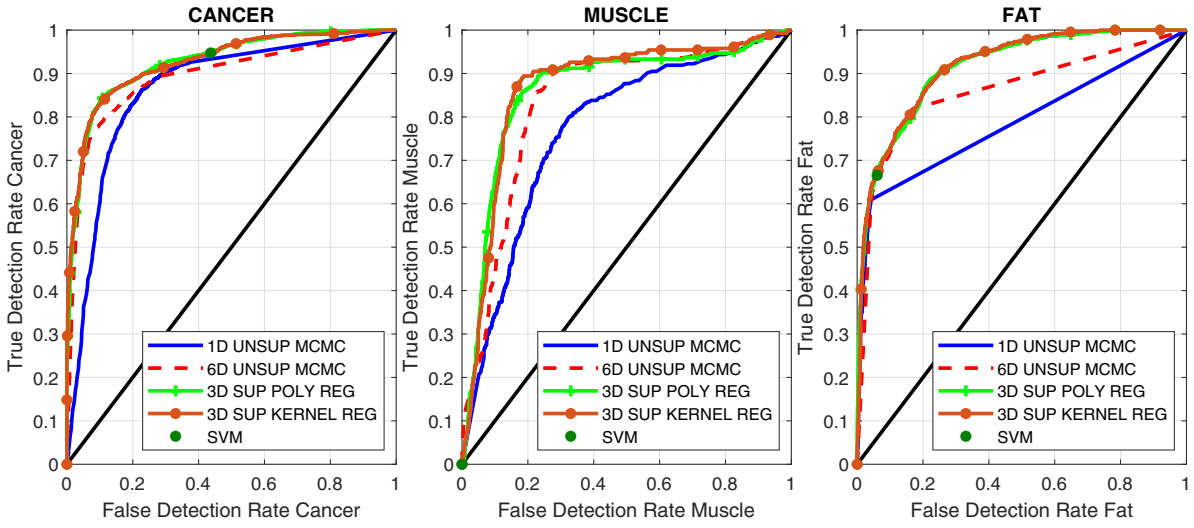

Fig. 10. Comparison of proposed classifiers vs. SVM.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the Oklahoma Animal Disease Diagnostic Laboratory and Oklahoma State University for handling the murine samples presented in this work. In addition, the authors would like to thank Dr. Rajaram and his students for their collaboration in handling the mice utilized in this study. This work was supported by the National Institutes of Health under Award No. R15CA208798. Additionally, it was supported in part by the National Science Foundation under Award 1711087.

Appendix A. ROC generation

An ROC curve illustrates the performance of a binary classifier. In a multi-class context, the classifier's performance is represented by multiple ROC curves with each of them corresponding to the detection of a given class against all the other classes, i.e. cancer vs. noncancer pixels in the THz image.

Let $P(y_n = k)$ denote the probability that the n-th pixel belongs to the k-th region. For a given threshold δ, the n-th pixel is classified as belonging to the k-th category if $P(y_n = k) \geqslant \delta$. Once δ is fixed, we can calculate the true detection rate and false detection rate by comparing the classification results with the morphed pathology, and this corresponds to one point on the ROC curve. A complete ROC curve can be obtained by varying the threshold value δ. In this paper, the ROC curve is generated by using the MATLAB function *perfcurve*, which utilizes the morphed pathology results as the ground truth information.

References

[1] A.C. Society, Cancer Facts & Figs. 2021, American Cancer Society, Atlanta, 2021.
[2] L.C. Elmore, J.A. Margenthaler, A tale of two operations: re-excision as a quality measure, Gland Surgery 8 (2019).
[3] Q. Mao, Y. Zhu, C. Lv, Y. Lu, X. Yan, S. Yan, J. Liu, Convolutional neural network model based on terahertz imaging for integrated circuit defect detections, Optics Express 28 (2020) 5000–5012.
[4] A. Golenkov, A. Shevchik-Shekera, M.Y. Kovbasa, I. Lysiuk, M. Vuichyk, S. Korinets, S. Bunchuk, S. Dukhnin, V. Reva, F. Sizov, THz linear array scanner in application to the real-time imaging and convolutional neural network recognition, Semiconductor Physics, Quantum Electronics & Optoelectronics 24 (2021) 90–99.
[5] Y. Shen, Y. Yin, B. Li, C. Zhao, G. Li, Detection of impurities in wheat using terahertz spectral imaging and convolutional neural networks, Computers and Electronics in Agriculture 181 (2021), 105931.
[6] T. Chavez, N. Vohra, J. Wu, K. Bailey, M. El-Shenawee, Breast cancer detection with low-dimension ordered orthogonal projection in terahertz imaging, IEEE Transactions on Terahertz Science and Technology (2019), 1–1.
[7] T. Chavez, N. Vohra, J. Wu, N. Rajaram, K. Bailey, M. El-Shenawee, Supervised statistical learning for cancer detection in dehydrated excised tissue with terahertz imaging, in: 2020 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting.
[8] W. Liu, R. Zhang, Y. Lu, R. She, K. Zhou, B. Fang, G. Wei, G. Li, Classification of terahertz pulsed signals from breast tissues using wavelet packet energy feature exaction and machine learning classifiers, in: C. Zhang, X.-C. Zhang, M. Tani (Eds.), Infrared, Millimeter-Wave, and Terahertz Technologies VI, volume 11196, International Society for Optics and Photonics, SPIE, 2019, pp. 15–22.
[9] F. Wahaia, I. Kašalynas, L. Minkevi-čius, C.C. Silva, A. Urbanowicz, G. Valušis, Terahertz spectroscopy and imaging for gastric cancer diagnosis, Journal of Spectral Imaging 9 (2020) a2.
[10] Y. Wang, Z. Sun, D. Xu, L. Wu, J. Chang, L. Tang, Z. Jiang, B. Jiang, G. Wang, T. Chen, H. Feng, J. Yao, A hybrid method based region of interest segmentation for continuous wave terahertz imaging, Journal of Physics D: Applied Physics 53 (2019), 095403.
[11] C. Hough, D.N. Purschke, C. Huang, L. Titova, O.V. Kovalchuk, B. Warkentin, F. A. Hegmann, Intense terahertz pulses inhibit ras signaling and other cancer-associated signaling pathways in human skin tissue models, Journal of Physics: Photonics (2021).
[12] N. Vohra, T. Chavez, J.R. Troncoso, N. Rajaram, J. Wu, P.N. Coan, T.A. Jackson, K. Bailey, M. El-Shenawee, Mammary tumors in Sprague Dawley rats induced by N-ethyl-N-nitrosourea for evaluating terahertz imaging of breast cancer, Journal of Medical Imaging 8 (2021) 1–17.
[13] F. Mendonça, S.S. Mostafa, F. Morgado-Dias, A.G. Ravelo-García, Cyclic alternating pattern estimation based on a probabilistic model over an EEG signal, Biomedical Signal Processing and Control 62 (2020), 102063.
[14] A. Anuragi, D.S. Sisodia, Empirical wavelet transform based automated alcoholism detecting using eeg signal features, Biomedical Signal Processing and Control 57 (2020), 101777.
[15] S. Helal, H. Sarieddeen, H. Dahrouj, T.Y. Al-Naffouri, M.S. Alouini, Signal processing and machine learning techniques for terahertz sensing: An overview, 2021.
[16] W. Liu, P. Zhao, Y. Shi, C. Liu, L. Zheng, Rapid determination of peroxide value of peanut oils during storage based on terahertz spectroscopy, Food Analytical Methods (2021) 1–9.
[17] P. Yang, D. Wang, W.-B. Zhao, L.-H. Fu, J.-L. Du, H. Su, Ensemble of kernel extreme learning machine based random forest classifiers for automatic heartbeat classification, Biomedical Signal Processing and Control 63 (2021), 102138.
[18] G. Gilanie, U.I. Bajwa, M.M. Waraich, M. Asghar, R. Kousar, A. Kashif, R.S. Aslam, M.M. Qasim, H. Rafique, Coronavirus (covid-19) detection from chest radiology images using convolutional neural networks, Biomedical Signal Processing and Control 66 (2021), 102490.
[19] A.K. Shukla, R.K. Pandey, R.B. Pachori, A fractional filter based efficient algorithm for retinal blood vessel segmentation, Biomedical Signal Processing and Control 59 (2020), 101883.
[20] N. Qi, Z. Zhang, Y. Xiang, Y. Yang, X. Liang, P. d. B. Harrington, Terahertz time-domain spectroscopy combined with support vector machines and partial least squares-discriminant analysis applied for the diagnosis of cervical carcinoma, Anal. Methods 7 (2015) 2333–2338.
[21] J.H. Albert, S. Chib, Bayesian analysis of binary and polychotomous response data, Journal of the American Statistical Association 88 (1993) 669–679.
[22] S.S. Chand, K.J.E. Walsh, Modeling seasonal tropical cyclone activity in the fiji region as a binary classification problem, Journal of Climate 25 (2012) 5057–5071.
[23] P. Milton, H. Coupland, E. Giorgi, S. Bhatt, Spatial analysis made easy with linear regression and kernels, Epidemics 29 (2019), 100362.
[24] D. Biswas, A. Gorey, G.C. Chen, S. Vasudevan, N. Sharma, P. Bhagat, S. Phatak, Empirical wavelet transform based photoacoustic spectral response technique for assessment of ex-vivo breast biopsy tissues, Biomedical Signal Processing and Control 51 (2019) 355–363.
[25] T. Chavez, T. Bowman, J. Wu, K. Bailey, M. El-Shenawee, Assessment of terahertz imaging for excised breast cancer tumors with image morphing, Journal of Infrared, Millimeter, and Terahertz Waves 39 (2018) 1283–1302.
[26] T. Bowman, T. Chavez, K. Khan, J. Wu, A. Chakraborty, N. Rajaram, K. Bailey, M. El-Shenawee, Pulsed terahertz imaging of breast cancer in freshly excised murine tumors, Journal of Biomedical Optics 23 (2018), 026004.
[27] N. Vohra, T. Bowman, K. Bailey, M. El-Shenawee, Terahertz imaging and characterization protocol for freshly excised breast cancer tumors, Journal of Visualized Experiments: JoVE (2020), e61007.
[28] T. Chavez, N. Vohra, M. El-Shenawee, K. Bailey, J. Wu, Source code for "multinomial probit regression for breast cancer detection in terahertz imaging, https://github.com/taxe10/Multinomial-Probit-Regression, 2021.
[29] T. Chavez, T. Bowman, J. Wu, M. El-Shenawee, K. Bailey, Cancer classification of freshly excised murine tumors with ordered orthogonal projection, in: 2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting, pp. 525–526.