5-2021

Statistical Machine Learning for Breast Cancer Detection with Terahertz Imaging

Tanny Andrea Chavez Esparza
University of Arkansas, Fayetteville

Statistical Machine Learning for Breast Cancer Detection with Terahertz Imaging

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Engineering with a concentration in Electrical Engineering

by

Tanny Andrea Chavez Esparza
Escuela Superior Politécnica del Litoral
Bachelor of Science in Electronics and Telecommunications Engineering, 2015
University of Arkansas
Master of Science in Electrical Engineering, 2018

May 2021
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Jingxian Wu, Ph.D.
Dissertation Director

Magda El-Shenawee, Ph.D.
Committee Member

Narasimhan Rajaram, Ph.D.
Committee Member

Jeff Dix, Ph.D.
Committee Member

Abstract

Breast conserving surgery (BCS) is a common breast cancer treatment option, in which the cancerous tissue is excised while leaving most of the healthy breast tissue intact. The lack of in-situ margin evaluation unfortunately results in a re-excision rate of 20-30% for this type of procedure. This study aims to design statistical and machine learning segmentation algorithms for the detection of breast cancer in BCS by using terahertz (THz) imaging. Given the material characterization properties of the non-ionizing radiation in the THz range, we intend to employ the responses from the THz system to identify healthy and cancerous breast tissue in BCS samples. In particular, this dissertation covers the description of four segmentation algorithms for the detection of breast cancer in THz imaging. We first explore the performance of one-dimensional (1D) Gaussian mixture and t-mixture models with Markov chain Monte Carlo (MCMC). Second, we propose a novel low-dimension ordered orthogonal projection (LOOP) algorithm for the dimension reduction of the THz information through a modified Gram-Schmidt process. Once the key features within the THz waveform have been detected by LOOP, the segmentation algorithm employs a multivariate Gaussian mixture model with MCMC and expectation maximization (EM). Third, we explore the spatial information of each pixel within the THz image through a Markov random field (MRF) approach. Finally, we introduce a supervised multinomial probit regression algorithm with polynomial and kernel data representations. For evaluation purposes, this study makes use of fresh and formalin-fixed paraffin-embedded (FFPE) heterogeneous human and mice tissue models for the quantitative assessment of the segmentation performance in terms of receiver operating characteristics (ROC) curves. Overall, the experimental results demonstrate that the proposed approaches represent a promising technique for tissue segmentation within THz images of freshly excised breast cancer samples.

Acknowledgments

This project was funded by the National Institutes of Health (NIH) award #R15CA208798, titled "Terahertz Imaging for Margin Assessment of Three Dimensional Breast Cancer Tumors" under the leadership of Dr. Magda El-Shenawee, PI, Dr. Narasimhan Rajaram, Co-PI, and Dr. Jingxian Wu, Co-PI, from March 2017 to February 2021. The work on unsupervised learning algorithms was partially funded by the National Science Foundation (NSF) under award #1711087. Additional funding was received from the University of Arkansas through a Distinguished Doctoral Fellowship from 2018-2021.

I would like to thank my advisor, Dr. Jingxian Wu, for his guidance and support along my doctoral studies. Beyond being thankful for the academic training, I appreciate his advice through the difficulties and obstacles encountered during the development of the algorithms summarized in this dissertation.

I wish to express my sincere thanks to my committee members: Dr. Magda El-Shenawee, Dr. Narasimhan Rajaram, and Dr. Jeff Dix, for their advice, feedback, and time to review the contents of this dissertation.

I would also like to thank the research team for their valuable efforts in the sample preparation process and hystopathology analysis. In particular, I am grateful to Nagma Vohra, PhD, candidate, for providing the raw terahertz imaging data from her experimental work using the terahertz imaging and spectroscopy system in Dr. El-Shenawee's lab at the University of Arkansas. I am grateful for Dr. Bailey's interpretations of the pathology images conducted by Ms. Nagma Vohra in the Histopathology Lab in the Biomedical Engineering Department at the University of Arkansas.

I am thankful to my friends near and far for their good wishes and unconditional support. Special thanks to my labmates at the Intelligent Information Processing lab at the University of Arkansas for their advice and motivation. I am forever grateful for our weekly meetings and the feedback I obtained from them, they truly helped me to keep moving forward.

I would like to thank my family for their long-distance, but ever-lasting love and support. In particular, to my parents and sister, who were there for me every step of the way. Finally, but not

least, I am extremely grateful to my fiancé, Sean, for his patience and encouragement throughout my doctoral studies.

Contents

List of Figures

List of Tables

List of Papers

Chapter 3:

The following papers have been edited and combined in this chapter:

- T. Chavez, N. Vohra, J. Wu, K. Bailey, and M. El-Shenawee, "Breast Cancer Detection with Low-Dimensional Ordered Orthogonal Projection in Terahertz Imaging," in *IEEE Transactions on Terahertz Science and Technology*, vol. 10, no. 2, pp. 176-189, March 2020, doi: 10.1109/TTHZ.2019.2962116.
- T. Chavez, N. Vohra, J. Wu, K. Bailey, and M. El-Shenawee, "Spatial Image Segmentation for Breast Cancer Detection in Terahertz Imaging," *2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting*, Montreal, QC, Canada, 2020, pp. 1157-1158, doi: 10.1109/IEEECONF35879.2020.9330445.

Chapter 4:

- T. Chavez, N. Vohra, K. Bailey, M. El-Shenawee, and J. Wu, "Supervised Bayesian Learning for Breast Cancer Detection in Terahertz Imaging", submitted for publication to *Biomedical Signal Processing and Control* and in review.

Chapter 1

Introduction

1.1. Motivation

Breast cancer corresponds to one of the most commonly diagnosed types of cancer with a higher incidence rate in women. According to [1], the U.S. estimates 284,200 new breast cancer cases in 2021, with 99.07% of them targeting women. Among different treatment options for this disease, breast-conserving surgery (BCS) represents a suitable procedure for early stage breast cancer patients, which is far less invasive than mastectomies.

BCS consists on excising the cancerous tissue from the breast surrounded by a small margin of healthy tissue, therefore leaving most of the non-cancerous tissue intact. Against common beliefs, extensive research has proven that BCS followed by radiation is as effective as mastectomy procedures, particularly for early stage breast cancer. Within the benefits of BCS we can mention: shorter recovery time, better aesthetics results, and fewer surgery complications. Despite these facts, a significant amount of BCS candidates choose mastectomies as part of their cancer treatment due to fear of cancer re-incidence and undergoing further medical procedures [2].

Considering that the chances of re-excision surgeries are associated with leaving cancerous tissue behind, the accurate detection of breast cancer cells in the operating room is a key missing component in BCS. To perform this detection in the state of art, surgeons send the excised tissue to a pathologist to verify the presence of negative margins in the sample, which confirms the successful extraction of the malign tissue. Although this procedure is effective, its main disadvantage is the long waiting time for the processing of the tissue, which takes around 10-15 days to be completed. As a consequence, BCS has a higher rate of re-excision surgeries due to positive margins after the pathology analysis, which varies from 20-30% in the literature [3]. Therefore, to decrease these re-excision rates, it is necessary to develop a breast cancer detection imaging benchmark for BCS that can be performed in the operating room.

1.2. Literature Review

Terahertz (THz) imaging is a non-ionizing radiation technique that has obtained potential characterization results for materials identification in different applications such as agriculture [4–6], concealed object detection [7–9], radar design [10–12], and biomedical imaging [13–28]. Within the latest, numerous studies have demonstrated the capacity of THz imaging for the detection of cancer due to the water's strong absorption properties in the THz frequency range [18–21, 29]. Since the water content and tissue density correspond to the key features for the detection of cancer, THz imaging is a strong candidate for the region segmentation of tumor samples in these applications [22]. Despite its potential, THz image segmentation approaches are still in an early stage of research and their detection rates vary across studies. Considering that doctors at the University of Massachusetts Medical School suggested cancer detection rates of no less than 90% [30], it is necessary to design a segmentation algorithm that complies with this requirement in order to establish a reliable clinical benchmark for THz imaging.

The literature presents a diverse array of segmentation approaches that rely on unsupervised and supervised learning methods. The selection of the learning technique is based in terms of the dataset size, ground truth availability, and prior knowledge on the application. For instance, unsupervised learning performs inference on the structure of the dataset without prior knowledge of its parameters distributions or ground truth. Therefore, unsupervised approaches are suitable for initial exploratory testing on the datasets. Alternatively, supervised learning techniques use part of the ground truth information to train their segmentation models. By performing this task, these models capture fundamental links among the THz data per pixel and their corresponding tissue regions, which are later used for discrimination purposes. Although supervised learning models have gained popularity for segmentation and classification problems, the requirement of large datasets, its complexity definition, and the bias-variance trade-off correspond to some of the obstacles for its complete adoption [31]. Therefore, this section introduces some commonly used segmentation techniques based on both unsupervised and supervised learning, and their results in

THz image segmentation.

Before introducing the details of existing segmentation algorithms, it is necessary to discuss the implementation of feature selection techniques and their impact on the performance of the segmentation process. According to Hughes phenomenon, the complexity of the segmentation process increases with the dimension size of the input information [32]. This means that while a high-dimensional input dataset provides plenty useful information, its dimension can negatively impact the performance of the segmentation algorithm. In contrast, a relatively small-dimensional input dataset will not provide enough data for the segmentation algorithm to work correctly. Hence, it is necessary to minimize the dimension size of the input data while minimizing the loss of information, such that the segmentation process can potentially obtain it's optimal performance. In THz research, some few studies rely on the utilization of physical characteristics within the reflected THz waveform, such as time-domain peak value [4, 5, 13], absorption coefficient, and refractive index [33]. Although the selection of single-feature characteristics minimizes the complexity of the segmentation process, the information captured within the THz waveform per pixel is not fully exploited. Alternatively, it is possible to automatically identify the fundamental features within the THz waveform through dimension reduction approaches, such as principal component analysis (PCA) [6, 19, 34, 35], partial least-squares discriminant analysis (PLS-DA) [36], non-negative matrix factorization (NMF) [35], and independent component analysis (ICA) [37]. Overall, the adoption of dimension reduction approaches can significantly reduce the impact of the curse of dimensionality while efficiently employing the key intrinsic features in the THz waveform per pixel.

Among unsupervised learning techniques, naive k-means is one of the most frequently used segmentation algorithms in THz imaging research. K-means is a hard clustering technique that utilizes the distance among each observation and k centroids to segment the data. Hence, this algorithm aims to minimize the squared error metric, which corresponds to the Eucledian distances between the observations assigned to a given cluster and its corresponding centroid [38]. Despite its simplicity, k-means is a non-deterministic polynomial-time (NP) hard problem and it is solved

heuristically. THz studies such as [18, 24] and [39] have utilized this technique for the segmentation of liver tumors and moth wings, respectively. These studies have shown that k-means is easy to implement, guarantees convergence, and can scale for large datasets. Unfortunately, its dependency on the initialization setup, the curse of dimensionality, and its hard-clustering results correspond to this method's major disadvantages. To address these problems, [39] presents a modified k-means clustering technique based on simple random sampling and feature weighting, which achieved promising detection rates in THz imaging of moth wings. Alternatively, fuzzy c-means (FCM) clustering introduces a membership weight parameter to the k-means segmentation problem, which results in a soft clustering classification approach. Since observations can belong to more than one cluster, this method outperforms k-means because it allows clusters to overlap. FCM segmentation has been tested on THz datasets in [18, 23, 24]. For instance, [23] implements a hybrid segmentation method for brain tumor detection in rats that consists in four stages: image denoising, FCM clustering, smooth contouring, and edge detection. This hybrid FCM method reported sensitivity and specificity rates of 84.5% and 97.7%, respectively, in a rat brain tumor sample.

On the other hand, supervised learning techniques have received acclaim for their outstanding performance in clustering and segmentation tasks. Within THz imaging research, methods such as support vector machine (SVM) [4, 6, 19, 25, 35, 36], region growing [40], support vector regression (SVR) [34, 41], k-nearest neighbors (KNN) [4, 19, 25, 35], and random forest (RF) [6, 25, 36] correspond to the most commonly used supervised learning algorithms. Among these methods, SVM aims to estimate the optimal discriminating hyperplane for a given training set, which can be implemented as a soft or hard clustering technique. Alternatively, KNN classifies a new observation by considering the labels of its k-nearest neighbors among the training dataset, where the final label is chosen by following a voting process. Finally, RF utilizes a decision tree approach based on bootstrapping theory, where testing observations are classified by popular vote among its decision trees. The study presented in [25] compared the performance of these three methods for the detection of traumatic brain injury in fresh rat samples. According to this work, KNN presented the best

precision results among the tested classifiers with rates of 87.5%, followed by RF with 84.2%, and SVM with 75%. The training and testing stages are handled by the leave-one-out cross validation (LOOCV) method and their classification results are presented as receiver operating characteristic (ROC) curves.

Within the supervised learning regimen, deep learning corresponds to a subset of techniques that implement hierarchical artificial neural networks (NN) for detection and estimation purposes. Inspired by the operation of neurons in the human brain, NN consist on a series of artificial neurons arranged in layers, such that each individual layer performs a simple task that collectively leads to a final complex discrimination result at the output of the network. The network architecture consists on an input layer, a fixed number of hidden layers, an output layer, and the connectors with their respective weights. Although the weights of the connectors are determined during the training phase, other aspects of the network should be established a priori, such as the number of hidden layers, and the activation function per layer. Therefore, these parameters are estimated using cross-validation approaches, which require extensive dataset availability. In THz research, [5, 7, 8] tested the performance of convolutional neural networks (CNN) for the detection of impurities in wheat, and concealed objects. According to [5], CNN has shown promising results for feature extraction while simplifying the data pre-processing procedure. Moreover, experimental results in different kinds of wheat samples showed that the testing accuracy reached a maximum of 97.83% for this discriminating method. Similarly, [6] developed a probabilistic NN (PNN) for the detection of liver cancer in THz imaging. Their proposed approach corresponds to a feed forward NN based on Bayesian theory by using a Gaussian kernel as the activation function. In this application, PNN achieved a maximum accuracy rate of 99.81% for the overall detection of the tumor. Considering both studies, NN is a potential candidate for the segmentation of THz images, but the need for large training datasets represents a major disadvantage for some applications.

In breast cancer research, the tissue collection process corresponds to an important aspect for the margin assessment of BCS samples. To establish a on-site clinical benchmark for THz imaging, it is necessary to quantitatively prove the accuracy of this technology by utilizing alternative

models that closely resembles the BCS human samples. Several THz studies have employed an assortment of tissue models, such as xenograft mice [13, 42], transgenic mice [14], Sprague-Dawley rats [15], phantom tissue [16], and human samples [17, 43]. It is important to clarify that the fresh human tissue utilized in THz research most commonly come from mastectomies and breast reduction surgeries, which are imaged within 24 hours of excision. Additionally, THz studies employ homogeneous and heterogeneous samples from these models, that can be imaged fresh or after dehydration. While some studies utilize homogeneous formalin-fixed paraffin-embedded (FFPE) BCS samples [19], the work presented in this dissertation focuses on heterogeneous fresh xenograft mice and human samples [43–45]. The utilization of these models represents an advantage due to the presence of different regions in the same sample, such cancer, fibro, collagen, fat, and others, which corresponds to a close representation of an actual BCS sample.

Overall, the literature presents an assortment of THz image segmentation approaches with performances that vary significantly across studies. These segmentation algorithms use unsupervised and supervised learning techniques depending on the size of the dataset, the ground truth availability, and the complexity of the segmentation task. Although further research in THz imaging segmentation is necessary, the presented studies unanimously prove the potential of THz imaging for material characterization, including breast cancer detection. Moreover, it is necessary to improve the image segmentation models to achieve the detection rates requirement for a BCS imaging benchmark based on THz technology.

1.3. Objectives

The aim of this dissertation is to implement novel image segmentation algorithms for the detection of breast cancer in BCS through THz imaging. The study objectives of each proposed algorithm are described below.

First, a one-dimensional Bayesian learning algorithm is implemented for the region segmentation in THz images of murine breast cancer samples. Considering that the prior knowledge of the dataset distribution is limited, we intend to analyze the performance of Gaussian and t-distributed

mixture models within an unsupervised Markov chain Monte Carlo (MCMC) estimation framework. Additionally, we examine the performance of this single-feature classifier by utilizing the power spectra and the amplitude peak of the THz waveform per pixel for fresh and FFPE tissue, respectively.

Second, a dimension reduction algorithm is developed to extract the most relevant intrinsic features within the THz waveform per pixel. For this purpose, the proposed algorithm implements a modified Gram-Schmidt process to estimate the orthonormal basis of these key features. Once this estimation process is completed, the features per pixel are obtained by projecting their THz waveforms into the subspace spanned by the orthonomal basis. In addition, a new Bayesian classifier is introduced, which employs a multivariate GMM approach with MCMC and EM. Overall, the performance of these and alternative algorithms are compared in fresh and FFPE human breast cancer samples.

Third, the spatial information per pixel is incorporated into the segmentation models. Considering that the reflected waveform per pixel is collected from an specific location in the sample, the correlation among neighboring pixels is higher when compared to those positioned at distant locations. Therefore, we aim to exploit the spatial information per pixel to account for this correlation in our previously implemented segmentation models through a Markov random field approach.

Finally, a supervised probit regression model is introduced for the region segmentation in THz images of freshly excised murine breast cancer tumors. The proposed algorithm considers two main aspects within this research project: the limited amount of samples available for training purposes, and the definition of the ground truth information due to tissue deformation during the pathology process. By using a probit regression approach with polynomial and kernel data representations, we reduce the number of estimation parameters, which decreases the amount of training observations that are required for convergence. Moreover, the proposed algorithm implements a novel reliability-based training selection process trough an unsupervised expectation maximization (EM) clustering approach.

1.4. Dissertation Outline

The rest of this dissertation is organized as follows:

Chapter 2: This chapter studies the implementation of Bayesian classifiers for the detection of breast cancer in murine samples. For this purpose, two different mixture models are adopted based on Gaussian and t-probability distributions. The proposed algorithms utilize a one-dimensional (1D) discriminating feature, such as the spectral power and the time-domain peak amplitude for fresh and FFPE tissue, respectively. Additionally, the parameter estimation process is performed through an unsupervised MCMC process. The performance of both algorithms is analyzed in murine samples with two and three regions.

Chapter 3: This chapter introduces a novel dimension reduction algorithm based on the estimation of orthonormal key features from a given THz dataset. Such algorithm implements a modified Gram-Schmidt process for the definition of its orthonormal basis, which contains the most relevant information within the THz pulses. By using this basis, the algorithm proceeds to project the THz waveform per pixel into the orthonormal basis to obtain their lower-dimensional representation. This chapter analyzes the performance of this dimension-reduction method for the segmentation of THz images through a Gaussian mixture model. Unlike chapter 2, the proposed classifier utilizes the magnitude and complex frequency-domain representation of the THz information, and its parameter estimation process is performed through unsupervised MCMC and EM algorithms. Finally, this chapter compares the performance of the proposed classifier with respect to alternative approaches, such as k-means, SVM, and PCA, in fresh and FFPE human breast cancer samples.

Additionally, this chapter builds on the algorithm presented in [43] by introducing the analysis of the neighborhood information within the probabilistic model definition. Based on the assumption that neighboring pixels have a higher probability to belong to the same region, the proposed algorithm aims to exploit the pixel location information through a Markov random field (MRF) framework. Hence, the overall segmentation process introduces a Gibbs prior to the region labels into the GMM classifier with EM. The performance of the proposed algorithm is then compared

with respect to the 1D MCMC classifier in fresh human breast cancer samples.

Chapter 4: This chapter analyzes the performance of supervised learning algorithms for breast cancer detection in THz images of fresh murine samples. For this purpose, we propose the implementation of probit regression-based segmentation approaches with two non-linear models, such as polynomial and kernel data representations. A key aspect of this approach consists on the small number of estimation parameters involved in these processes, which results on a small training set size. It is important to consider that the ground truth information utilized during the training process is collected from the histopathology analysis of the samples after dehydration, while the final evaluation of the segmentation algorithm is observed within the THz data collected when the tissue was still fresh. Hence, the proposed methodology presents a training selection algorithm that evaluates the reliability of the training observations by using an unsupervised EM approach.

Chapter 5: This chapter introduces the conclusion remarks and major contributions of this work. Additionally, a list of potential research subjects is presented in this chapter for future consideration.

References

[1] A. C. Society, *Cancer Facts & Figures 2021*. Atlanta: American Cancer Society, 2021, Accessed: Feb. 18, 2021. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf

[2] American Cancer Society, *Cancer Facts & Figures 2019-2020*. Atlanta: American Cancer Society, Inc. 2019, 2020, Accessed: Feb. 14, 2020. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/

[3] L. C. Elmore and J. A. Margenthaler, "A tale of two operations: re-excision as a quality measure," *Gland Surgery*, vol. 8, no. 6, 2019, Accessed: Feb. 14, 2020. [Online]. Available: http://gs.amegroups.com/article/view/32941

[4] B. Li, D. Zhang, and Y. Shen, "Study on terahertz spectrum analysis and recognition modeling of common agricultural diseases," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 243, p. 118820, 2020. doi: https://doi.org/10.1016/j.saa.2020.118820. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138614252030799X

[5] Y. Shen, Y. Yin, B. Li, C. Zhao, and G. Li, "Detection of impurities in wheat using terahertz spectral imaging and convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 181, p. 105931, 2021. doi: https://doi.org/10.1016/j.compag.2020.105931. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169920331367

[6] W. Liu, C. Liu, X. Hu, J. Yang, and L. Zheng, "Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics," *Food Chemistry*, vol. 210, pp. 415 – 421, 2016. doi: https://doi.org/10.1016/j.foodchem.2016.04.117. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0308814616306458

[7] M. Kovbasa, A. Golenkov, and F. Sizov, "Neural network application to the postal terahertz scanner for automated detection of concealed items," in *2020 IEEE Ukrainian Microwave Week (UkrMW)*, 2020, pp. 870–873. doi: 10.1109/UkrMW49653.2020.9252706

[8] H. Feng, D. An, H. Tu, W. Bu, W. Wang, Y. Zhang, H. Zhang, X. Meng, W. Wei, B. Gao, and S. Wu, "A passive video-rate terahertz human body imager with real-time calibration for security applications," *Applied Physics B: Lasers and Optics*, vol. 126, no. 8, p. 143, Aug. 2020. doi: 10.1007/s00340-020-07496-3

[9] Y. Cheng, Y. Wang, Y. Niu, and Z. Zhao, "Concealed object enhancement using multi-polarization information for passive millimeter and terahertz wave security screening," *Opt. Express*, vol. 28, no. 5, pp. 6350–6366, Mar 2020. doi: 10.1364/OE.384029. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-28-5-6350

[10] Z. Ou, J. Wu, H. Geng, X. Deng, and X. Zheng, "Confocal terahertz SAR imaging of hidden objects through rough-surface scattering," *Opt. Express*, vol. 28, no. 8, pp. 12 405–12 415, Apr 2020. doi: 10.1364/OE.388392. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-28-8-12405

[11] G. Wang, F. Qi, Z. Liu, C. Liu, C. Xing, and W. Ning, "Comparison between back projection algorithm and range migration algorithm in terahertz imaging," *IEEE Access*, vol. 8, pp. 18 772–18 777, 2020. doi: 10.1109/ACCESS.2020.2968085

[12] Y. Li, Q. Wu, J. Wu, P. Li, Q. Zheng, and L. Ding, "Estimation of high-frequency vibration parameters for terahertz SAR imaging based on FrFT with combination of QML and RANSAC," *IEEE Access*, vol. 9, pp. 5485–5496, 2021. doi: 10.1109/ACCESS.2020.3047856

[13] T. Bowman, T. Chavez, K. Khan, J. Wu, A. Chakraborty, N. Rajaram, K. Bailey, and M. El-Shenawee, "Pulsed terahertz imaging of breast cancer in freshly excised murine tumors," *Journal of Biomedical Optics*, vol. 23, no. 2, p. 026004, 2018. doi: 10.1117/1.JBO.23.2.026004.

[14] N. Vohra, T. Bowman, P. M. Diaz, N. Rajaram, K. Bailey, and M. El-Shenawee, "Pulsed terahertz reflection imaging of tumors in a spontaneous model of breast cancer," *Biomedical Physics & Engineering Express*, vol. 4, no. 6, p. 065025, oct 2018. doi: 10.1088/2057-1976/aae699. [Online]. Available: https://doi.org/10.1088/2057-1976/aae699

[15] N. Vohra, T. Chavez, J. R. Troncoso, N. Rajaram, J. Wu, P. N. Coan, T. A. Jackson, K. Bailey, and 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*, vol. 8, no. 2, pp. 1 – 17, 2021. doi: 10.1117/1.JMI.8.2.023504. [Online]. Available: https://doi.org/10.1117/1.JMI.8.2.023504

[16] T. Bowman, A. Walter, O. Shenderova, N. Nunn, G. McGuire, and M. El-Shenawee, "A phantom study of terahertz spectroscopy and imaging of micro-and nano-diamonds and nano-onions as contrast agents for breast cancer," *Biomedical physics & engineering express*, vol. 3, no. 5, p. 055001, 2017.

[17] T. Bowman, N. Vohra, K. Bailey, and M. O. El-Shenawee, "Terahertz tomographic imaging of freshly excised human breast tissues," *Journal of Medical Imaging*, vol. 6, no. 2, pp. 1 – 13, 2019. doi: 10.1117/1.JMI.6.2.023501. [Online]. Available: https://doi.org/10.1117/1.JMI.6.2.023501

[18] H. Liu, Z. Zhang, and C. Zhang, "Enhance the contrast for the terahertz pulse parametric imaging," in *Infrared, Millimeter-Wave, and Terahertz Technologies V*, C. Zhang, X.-C. Zhang, and M. Tani, Eds., vol. 10826, International Society for Optics and Photonics. SPIE, 2018, pp. 196 – 201. doi: 10.1117/12.2500946. [Online]. Available: https://doi.org/10.1117/12.2500946

[19] W. Liu, R. Zhang, Y. Ling, H. Tang, R. She, G. Wei, X. Gong, and Y. Lu, "Automatic recognition of breast invasive ductal carcinoma based on terahertz spectroscopy

with wavelet packet transform and machine learning," *Biomed. Opt. Express*, vol. 11, no. 2, pp. 971–981, Feb 2020. doi: 10.1364/BOE.381623. [Online]. Available: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-11-2-971

[20] K. Okada, K. Serita, Q. Cassar, H. Murakami, G. MacGrogan, J.-P. Guillet, P. Mounaix, and M. Tonouchi, "Terahertz near-field microscopy of ductal carcinoma in situ (DCIS) of the breast," *Journal of Physics: Photonics*, vol. 2, no. 4, p. 044008, oct 2020. doi: 10.1088/2515-7647/abbcda. [Online]. Available: https://doi.org/10.1088/2515-7647/abbcda

[21] R. Grigorev, A. Kuzikova, P. Demchenko, A. Senyuk, A. Svechkova, A. Khamid, A. Zakharenko, and M. Khodzitskiy, "Investigation of fresh gastric normal and cancer tissues using terahertz time-domain spectroscopy," *Materials*, vol. 13, no. 1, 2020. doi: 10.3390/ma13010085. [Online]. Available: https://www.mdpi.com/1996-1944/13/1/85

[22] T. C. Bowman, M. El-Shenawee, and L. K. Campbell, "Terahertz imaging of excised breast tumor tissue on paraffin sections," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 5, pp. 2088–2097, 2015. doi: 10.1109/TAP.2015.2406893

[23] Y. Wang, Z. Sun, D. Xu, L. Wu, J. Chang, L. Tang, Z. Jiang, B. Jiang, G. Wang, T. Chen, H. Feng, and J. Yao, "A hybrid method based region of interest segmentation for continuous wave terahertz imaging," *Journal of Physics D: Applied Physics*, vol. 53, no. 9, p. 095403, dec 2019. doi: 10.1088/1361-6463/ab58b6. [Online]. Available: https://doi.org/10.1088%2F1361-6463%2Fab58b6

[24] H. Liu, Z. Zhang, and C. Zhang, "Enhance the contrast for the terahertz pulse parametric imaging," in *Infrared, Millimeter-Wave, and Terahertz Technologies V*, C. Zhang, X.-C. Zhang, and M. Tani, Eds., vol. 10826, International Society for Optics and Photonics. SPIE, 2018, pp. 196 – 201. doi: 10.1117/12.2500946. [Online]. Available: https://doi.org/10.1117/12.2500946

[25] J. Shi, Y. Wang, T. Chen, D. Xu, H. Zhao, L. Chen, C. Yan, L. Tang, Y. He, H. Feng, and J. Yao, "Automatic evaluation of traumatic brain injury based on terahertz imaging with machine learning," *Opt. Express*, vol. 26, no. 5, pp. 6371–6381, Mar 2018. doi: 10.1364/OE.26.006371. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-26-5-6371

[26] C. Hough, D. N. Purschke, C. Huang, L. Titova, O. V. Kovalchuk, B. Warkentin, and 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. [Online]. Available: http://iopscience.iop.org/article/10.1088/2515-7647/abf742

[27] Z. Yu and L. Zhang, "Research progress and prospects of the biological effects of terahertz radiation," *Journal of the Third Military Medical University*, vol. 42, no. 23, pp. 2259–2266, 2020.

[28] F. Behague, V. Calero, A. Coste, A. Godet, M. Suarez, G. Gaborit, L. Duvillaret, F. I. Baida, M.-P. Bernal, and N. Courjal, "Minimally invasive optical sensors for microwave-electric-field exposure measurements," *Journal of Optical Microsystems*, vol. 1, no. 2, pp. 1 – 17, 2021. doi: 10.1117/1.JOM.1.2.020902. [Online]. Available: https://doi.org/10.1117/1.JOM.1.2.020902

[29] X. Yang, X. Zhao, K. Yang, Y. Liu, Y. Liu, W. Fu, and Y. Luo, "Biomedical applications of terahertz spectroscopy and imaging," *Trends in Biotechnology*, vol. 34, no. 10, pp. 810 – 824, 2016. doi: https://doi.org/10.1016/j.tibtech.2016.04.008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167779916300270

[30] B. St. Peter, S. Yngvesson, P. Siqueira, P. Kelly, A. Khan, S. Glick, and A. Karellas, "Development and testing of a single frequency terahertz imaging system for breast cancer detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 785–797, July 2013. doi: 10.1109/JBHI.2013.2267351

[31] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised machine learning: A brief primer," *Behavior Therapy*, vol. 51, no. 5, pp. 675–687, 2020. doi: https://doi.org/10.1016/j.beth.2020.05.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005789420300678

[32] R. Bellman, R. Bellman, and K. M. R. Collection, *Adaptive Control Processes: A Guided Tour*, ser. Princeton Legacy Library. Princeton University Press, 1961. ISBN 9780691079011. [Online]. Available: https://books.google.com/books?id=POAmAAAAMAAJ

[33] B. Li, X. Zhao, Y. Zhang, S. Zhang, and B. Luo, "Prediction and monitoring of leaf water content in soybean plants using terahertz time-domain spectroscopy," *Computers and Electronics in Agriculture*, vol. 170, p. 105239, 2020. doi: https://doi.org/10.1016/j.compag.2020.105239. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169919317168

[34] Y. Sun, P. Du, X. Lu, P. Xie, Z. Qian, S. Fan, and Z. Zhu, "Quantitative characterization of bovine serum albumin thin-films using terahertz spectroscopy and machine learning methods," *Biomed. Opt. Express*, vol. 9, no. 7, pp. 2917–2929, Jul 2018. doi: 10.1364/BOE.9.002917. [Online]. Available: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-9-7-2917

[35] S. Helal, H. Sarieddeen, H. Dahrouj, T. Y. Al-Naffouri, and M. S. Alouini, "Signal processing and machine learning techniques for terahertz sensing: An overview," *arXiv preprint arXiv:2104.06309*, 2021.

[36] J. Zhang, Y. Yang, X. Feng, H. Xu, J. Chen, and Y. He, "Identification of bacterial blight resistant rice seeds using terahertz imaging and hyperspectral imaging combined with convolutional neural network," *Frontiers in Plant Science*, vol. 11, p. 821, 2020. doi: 10.3389/fpls.2020.00821. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpls.2020.00821

[37] J. Xu, H. Wang, Y. Duan, Y. He, S. Chen, and Z. Zhang, "Terahertz imaging and vibrothermography for impact response in carbon fiber reinforced plastics," *Infrared Physics & Technology*, vol. 109, p. 103413, 2020. doi: https://doi.org/10.1016/j.infrared.2020.103413. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1350449520304618

[38] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, March 1982. doi: 10.1109/TIT.1982.1056489

[39] M. W. Ayech and D. Ziou, "Terahertz image segmentation using k-means clustering based on weighted feature learning and random pixel sampling," *Neurocomputing*, vol. 175, pp. 243 – 264, 2016. doi: https://doi.org/10.1016/j.neucom.2015.10.056. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215015155

[40] T. Bowman, Y. Wu, J. Gauch, L. K. Campbell, and M. El-Shenawee, "Terahertz Imaging of Three-Dimensional Dehydrated Breast Cancer Tumors," *Journal of Infrared*, vol. 38, no. 6, pp. 766–786, Jun. 2017. doi: 10.1007/s10762-017-0377-y

[41] Y. Peng, C. Shi, M. Xu, T. Kou, X. Wu, B. Song, H. Ma, S. Guo, L. Liu, and Y. Zhu, "Qualitative and quantitative identification of components in mixture by terahertz spectroscopy," *IEEE Transactions on Terahertz Science and Technology*, vol. 8, no. 6, pp. 696–701, 2018.

[42] T. Chavez, T. Bowman, J. Wu, K. Bailey, and M. El-Shenawee, "Assessment of terahertz imaging for excised breast cancer tumors with image morphing," *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 39, no. 12, pp. 1283–1302, Dec 2018. doi: 10.1007/s10762-018-0529-8. [Online]. Available: https://doi.org/10.1007/s10762-018-0529-8

[43] T. Chavez, N. Vohra, J. Wu, K. Bailey, and M. El-Shenawee, "Breast cancer detection with low-dimension ordered orthogonal projection in terahertz imaging," *IEEE Transactions on Terahertz Science and Technology*, pp. 1–1, 2019. doi: 10.1109/TTHZ.2019.2962116

[44] T. Chavez, N. Vohra, J. Wu, K. Bailey, and M. El-Shenawee, "Spatial image segmentation for breast cancer detection in terahertz imaging," in *2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting*, 2020, pp. 1157–1158. doi: 10.1109/IEEECONF35879.2020.9330445

[45] T. Chavez, N. Vohra, K. Bailey, M. El-Shenawee, and J. Wu, "Supervised bayesian learning for breast cancer detection in terahertz imaging," submitted for publication.

Chapter 2

Unsupervised Bayesian Learning for Cancer Detection with Terahertz Imaging

2.1. Abstract

This chapter develops unsupervised Bayesian learning algorithms for breast cancer detection in Terahertz (THz) imaging of freshly excised murine tumors. Unlike most existing works with deterministic detection methods, we adopt a probabilistic learning approach that can iteratively calculate the probability each pixel in a THz image belonging to different types of tissues, such as cancer, fat, muscle, fibrous tissue, etc. Such a probabilistic approach produces important reliability information about the detection results that are not available in conventional methods. Specifically, under a Bayesian framework, a finite mixture model is used to represent the probability distributions of the intensities of pixels in the THz image, with each component in the mixture model corresponding to one tissue type. The prevalence of a specific type of tissue in a pixel can be represented through the weights of corresponding component to be learned through the data, without the need of labeled training data. Two different mixture models, Gaussian mixture and t-mixture models, are employed in the analysis. The empirical posterior distributions of parameters from both models are estimated by using a Markov chain Monte Carlo (MCMC) technique with Gibbs sampling. The performance of the algorithms is evaluated by comparing the detection results to their corresponding pathology results, and experiment results demonstrate the proposed algorithm can classify different tissue types with high accuracy. Overall, THz imaging shows good qualitative comparison to pathology.

2.2. Introduction

Breast cancer is the second most commonly diagnosed type of cancer and it has the second highest cancer death rate for women in the United States [1]. When breast cancer is detected in its early stage, breast-conserving surgery is a feasible treatment option in which the breast cancer tumors

are excised entirely while leaving as much healthy breast tissue as possible. For this procedure to be successful, it is necessary to accurately distinguish between healthy and cancer tissues. This motivates us to develop new and efficient cancer detection technologies by using THz imaging of freshly excised tissues, which can be performed in the operation room to assess the operation results and avoid multiple surgeries days later.

The potential of THz imaging for breast cancer detection has been studied for both freshly excised tissue and formalin-fixed, paraffin-embedded (FFPE) tissue using time and frequency domain techniques to distinguish between cancer and healthy regions within the tumor [2–6]. Some means of observing contrast in THz imaging include the peak-to-peak ratio of the THz pulse [7], obtaining a multispectral integration of the Fourier transform [2], or calculating optical properties from the reflected signal [4]. Recently there has been growing interest in applying machine learning and statistical learning methods for cancer tissue detection with THz imaging, such as support vector machine (SVM) [8–10], principal component analysis (PCA) [8], decision tree analysis [9, 10], and neural network analysis (NNA) [9, 10]. The results in [8] demonstrated that the combination of data reduction with PCA and classification with SVM can achieve accurate cancer detection. The performance of SVM, decision trees, and NNA was compared for the detection of colon cancer [9] and gastric cancer [10]. According to the experimental results in [9], the application of NNA and PCA procedures for the THz classification achieved the best outcome to distinguish normal from abnormal (pre-cancerous and cancer) tissue, while the decision tree technique achieved the best classification for normal and pre-cancerous tissue. The results in [10] highlight that, when there is sufficient amount of training data, NNA can effectively classify normal and abnormal tissues in THz imaging of gastric cancer tumors.

All above works were based on supervised learning, where an extensive amount of training is required before the detection. However, different tumors might have different mixtures of cancer, fat, muscle, fibrous tissue, etc; thus the training results obtained from one group of tissue samples might not work well on other tissue samples. In addition, most existing works assume a deterministic detection approach, which treats each pixel in the THz image as coming exclusively

from either cancer or normal tissue, while in reality the signal on a THz image pixel might contain contribution from a mixture of different types of tissues due to the heterogeneity of breast tumors.

We propose to address the above problems by developing an unsupervised Bayesian learning framework with finite mixture models, which can identify, learn, and adapt to the properties of different types of tissues without requiring a training phase. Finite mixture analysis is a robust technique for statistical data modeling and has been widely applied on image processing routines such as image segmentation [11] and image reconstruction [12]. This modeling tool assumes that each element within the data follows a certain probability distribution from a set of possible distributions, and it infers the parameters of these distributions to finally cluster the elements in the data [13]. Mixture analysis has been previously applied in unsupervised spectral unmixing for hyperspectral analysis to estimate the endmembers and their corresponding abundance [14, 15]. Both [14, 15] adopt a Bayesian framework in which the posterior distribution of their parameters is computed through a Markov chain Monte Carlo [16] algorithm with Gibbs sampling. Alternative applications of mixture analysis in THz imaging classification has been documented in [17] for successful object identification (e.g. hazardous object scanning at airports) through temperature estimation.

The objective of this chapter is to develop unsupervised Bayesian learning algorithms that can detect different regions within a frehsly excised murine breast cancer tumor using its THz image, without the need of a training phase. A finite mixture model is used to represent the statistical distributions of the intensities of pixels, with each component in the mixture model corresponding to one tissue type. The mixture model will be implemented here on the images of freshly excised murine tumors [14, 15]. The proposed algorithms are probabilistic in nature, in that they will calculate the probability that each pixel in the THz image belonging to different tissue types. Such a probabilistic approach can quantify the uncertainty regarding the detection results that is not available in the deterministic approaches.

Adopting a Bayesian framework, the proposed algorithms are developed by following the MCMC scheme [16] with Gibbs sampling [18] and [19]. Two different mixture models, Gaus-

sian mixture model and t-mixture model, are studied in this chapter. The results of both models are evaluated in fresh and FFPE tissues of murine tumors using a pixel-by-pixel comparison with their corresponding pathology images. Following some of the authors' previous work in [3, 4], where the accuracy of the tissue classification was assessed qualitatively, this chapter presents a quantitative evaluation of the results similar to [2, 20]. Therefore, the results are presented in the form of receiver operation characteristic (ROC) curves, which show the true positive ratio as the function of false positive ratio. Experimental results indicate that the proposed algorithms can achieve good accuracy in identifying different regions within a tumor.

The chapter is organized as follows. Section 2.3 presents the framework of the unsupervised Bayesian learning algorithm. The detailed implementations of the algorithm with the Gaussian and t-mixture models are developed in Sections 2.4 and 2.5, respectively. Section 2.6 presents the experimental results, and Section 2.7 concludes the chapter.

2.3. Unsupervised Learning Through Bayesian Mixture Model

2.3.1. Bayesian Mixture Model

In this section, we present an unsupervised learning method of the THz image based on Bayesian mixture models. The objective is to classify each pixel in the THz image into one of several categories based on the statistical properties of the THz signal, without a training phase.

Define the pixel intensity vector $\mathbf{x} = [x_1, \cdots, x_N]^T \in \mathcal{R}_+^{N\times 1}$, where $\mathcal{R}_+$ is the set of non-negative real numbers. The variable x_i represents the summarized intensity (e.g. frequency domain integration for fresh tissue and time domain peak for FFPE tissue) at the i-th pixel, and its value is directly related to the physical properties of the tissue corresponding to that pixel.

In the proposed unsupervised Bayesian learning method, we assume that the summarized intensity has a multi-modal probability distribution, with each mode corresponding to one possible category. Given the summarized intensity of each pixel, the proposed algorithm calculates the multi-modal probability distribution using Bayesian inference and computes the posterior probabilities of each pixel belonging to different categories. During the development of the algorithm,

it is assumed that the number of modes is known, and the intensities of pixels corresponding to any particular tissue type follows an unimodal distribution. The final classification is performed by selecting the category with the highest probability for each pixel in the THz image.

Assume each pixel can be classified into one of k categories. Define a sequence of independent latent variable $z_i \in \{1, 2, ..., k\}$, for $i = 1, \cdots, N$, where $z_i = j$ means that the i-th pixel belongs to the j-th category, for $j = 1, \cdots, k$. The latent variable follows a multinomial distribution, with the probability mass function (PMF) represented as

$$\pi(z_i = j) = q_j,$$

where q_j represents the prior probability of any pixel belonging to the j-th category.

In Bayesian inference, the prior probability vector $\mathbf{q} = [q_1, \cdots, q_k] \in \mathcal{L}^{k\times 1}$ with $\mathcal{L} = [0, 1]$ is unknown and is usually assumed to be a random vector that follows the Dirichlet distribution, i.e., $\pi(\mathbf{q}) = \text{Dir}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_k]^T \in \mathcal{R}^{k\times 1}$ represents the parameter of the Dirichlet distribution.

We employ a multi-modal mixture model for the probability density function (pdf) of x_i, and it can be represented as

$$f(x_i|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_k, \mathbf{q}) = \sum_{j=1}^{k} q_j f(x_i|z_i = j, \boldsymbol{\theta}_j) \tag{2.1}$$

where $f(x_i|z_i = j, \boldsymbol{\theta}_j)$ is the likelihood function of x_i given pixel i is in the j-th category, and $\boldsymbol{\theta}_j = [\theta_{j1}, \cdots, \theta_{jp}] \in \mathcal{R}^{p\times 1}$ is used to represent the p unknown parameters of $f(x_i|z_i = j, \boldsymbol{\theta}_j)$. In Bayesian inference, the parameters $\boldsymbol{\theta}_j$ are assumed to be unknown random vectors with prior distribution $\pi(\boldsymbol{\theta}_j)$.

It is well known that the optimum classifier that can minimize the classification error is the maximum a posterior probability (MAP) classifier, which maximizes the posterior probability of

z_i given the signal intensity as

$$\hat{z}_i = \underset{j \in \{1, \cdots, k\}}{\operatorname{argmax}} \ \Pr(z_i = j | x_i) \tag{2.2}$$

where $\hat{z}_i$ is the classification result, and the posteriori probability can be calculated as

$$\Pr(z_i = j | x_i) = \frac{\int_{\boldsymbol{\theta}_j} \int_{q_j} f(x_i | z_i = j, \boldsymbol{\theta}_j) q_j \pi(\boldsymbol{\theta}_j) \pi(q_j) d\boldsymbol{\theta}_j dq_j}{f(x_i)}. \tag{2.3}$$

The posterior probability defined in (2.3) requires the knowledge of the prior distributions $\pi(\boldsymbol{\theta}_j)$ and $\pi(q_j)$, whose parameters were assumed to be known. In addition, the calculation of the posterior probability requires multi-level integrations with respect to the multi-dimensional parameter vector $\boldsymbol{\theta}_j$ and q_j, which are usually difficult to carry out either analytically or numerically.

2.3.2. MCMC with Gibbs Sampling

We propose to solve the problem by employing MCMC, which can obtain a numerical approximation of $\Pr(z_i = j | \mathbf{x})$ by iteratively taking samples from the joint distribution $f(\mathbf{z}, \{\boldsymbol{\theta}_j\}_{j=1}^k, \mathbf{q} | \mathbf{x})$. Assume that for a given $\mathbf{x}$, we have T samples taken from the joint distribution and denote them as $\{\mathbf{z}^{(it)}\}_{it=1}^T$, $\{\boldsymbol{\theta}_j^{(it)}\}_{it=1}^T$, for $j = 1, \cdots, k$, and $\{\mathbf{q}^{(it)}\}_{it=1}^T$. Then based on the law of large numbers, as $T \to \infty$, we have

$$\Pr(z_i = j | \mathbf{x}) = \lim_{T \to \infty} \frac{1}{T} \sum_{it=1}^{T} \mathcal{I}(z_i^{(it)} = j), \tag{2.4}$$

where $\mathcal{I}(\mathcal{E})$ is an indicator function defined as $\mathcal{I}(\mathcal{E}) = 1$ if $\mathcal{E}$ is true and 0 otherwise.

The basic idea of MCMC with Gibbs sampling is to iteratively take samples based on the posterior distributions of different variables conditioned on previously taken samples, thus effectively tackle the "curse-of-dimensionality" problem that plagued high dimension sampling. Details of Gibbs sampling under the Gaussian-mixture model and t-mixture model will be presented in the next two sections, respectively.

In order to initialize the iterative sampling process, the values of the unknown variables and parameters need to be initialized. We propose to initialize the values by using results from K-means clustering.

With the K-means algorithm, the vector $\mathbf{x}$ is classified into k categories. Consider the set of pixels that correspond to the j-th component as $\mathcal{S}_j^{(0)} = \{i : z_i = j\}$ with cardinality $n_j^{(0)} = |\mathcal{S}_j^{(0)}|$. Initialize $\mathbf{z}^{(0)}$ by assigning $z_i^{(0)} = j$ if $i \in \mathcal{S}_j$. Then the vector $\mathbf{q}^{(0)}$ can be initialized as

$$q_j^{(0)} = \frac{n_j^{(0)}}{N}, \quad j = 1, \cdots, k. \tag{2.5}$$

Define a vector $\mathbf{x}_j = [x_i]_{i \in \mathcal{S}_j} \in \mathcal{R}_+^{n_j \times 1}$, which contains x_i corresponding to all pixels labeled as $z_i = j$. The unknown parameters $\boldsymbol{\theta}_j$ can then be estimated from $\mathbf{x}_j$ by using maximum likelihood estimation. In this chapter we will consider two different models: Gaussian mixture model and t-mixture model, and both require the parameters of mean μ_j and variance σ_j^2. The unbiased estimation of μ_j and σ_j^2 can be written as

$$\mu_j^{(0)} = \frac{1}{n_j^{(0)}} \sum_{i \in \mathcal{S}_j} x_i \tag{2.6}$$

$$(\sigma_j^2)^{(0)} = \frac{1}{n_j^{(0)} - 1} \sum_{i \in \mathcal{S}_j} \left(x_i - \mu_j^{(0)} \right)^2 \tag{2.7}$$

The initialization of parameters other than mean or variance is model dependent and will be discussed for their specific models in the subsequent sections.

2.3.3. Unsupervised Bayesian Learning with MCMC

With the initial values of $\mathbf{z}^{(0)}$, $\mathbf{q}^{(0)}$, and $\boldsymbol{\theta}_j^{(0)}$, we summarize the outline of the unsupervised Bayesian learning algorithm with MCMC in algorithm 1.

In Gibbs sampling, the generated samples at the beginning of the sampling process usually do not represent the actual joint distribution, therefore we usually discard the first $T_0 - 1$ samples during the evaluation process as shown in (2.8).

Algorithm 1 Unsupervised Bayesian Learning with MCMC

Input: Pixel intensity vector $\mathbf{x}$.
Initialize: Obtain $\mathbf{z}^{(0)}$, $\mathbf{q}^{(0)}$, and $\{\boldsymbol{\theta}_j^{(0)}\}_{j=1}^k$ by using results from K-means clustering of $\mathbf{x}$.
Gibbs Sampling: Draw T samples $\mathbf{z}^{(it)}$, $\mathbf{q}^{(it)}$, and $\{\boldsymbol{\theta}_j^{(it)}\}_{j=1}^k$, for $it = 1, \cdots, T$, by using their respective posterior distributions. (Details are in Sections 2.4 and 2.5.)
Evaluation: Calculate the posterior probability $\Pr(z_i = j|\mathbf{x})$ as,

$$\Pr(z_i = j|\mathbf{x}) = \frac{1}{T}\sum_{it=T_0}^{T} \mathcal{I}(z_i^{(it)} = j), \tag{2.8}$$

for $i = 1, \cdots, N$, and $j = 1, \cdots, k$.
Output:

$$\hat{z}_i = \underset{j\in\{1,\cdots,k\}}{\operatorname{argmax}} \ \Pr(z_i = j|\mathbf{x}) \tag{2.9}$$

It is important to highlight that since there is not a natural ordering between mixture components, it is necessary to label them for their posterior identification (cancer, muscle, etc.). For our algorithm, we labeled the components according to the increasing order of their means following the labeling criterion described in [21], meaning that the components are ordered with the assumption that fat tissue will have the lowest reflected signal, followed by fibrous or muscle tissue and finally cancer tissue.

Details of Gibbs sampling for Gaussian-mixture model and t-mixture model are discussed in the next two sections, respectively.

2.4. Gibbs Sampling with Gaussian Mixture Model

In this section, we adopt the Gaussian mixture model for the unsupervised Bayesian learning algorithm, and develop the detailed procedures used by Gibbs sampling. In the Gaussian mixture model, the likelihood function of x_i conditioned on $z_i = j$ follows a Gaussian distribution as

$$f(x_i|z_i = j, \mu_j, \sigma_j^2) = \mathcal{N}(\mu_j, \sigma_j^2), j = 1, \cdots, k, \tag{2.10}$$

where μ_j and σ_j^2 represent the mean and the variance of the j-th component, respectively, and they are treated as unknown random variables under the Bayesian setting. Thus for the Gaussian mixture model, the unknown model parameters are $\boldsymbol{\theta}_j = [\mu_j, \sigma_j^2]^T$, for $j = 1, \cdots, k$.

The priors for the mean and variance are represented as:

$$\pi(\mu_j) = \mathcal{N}(\mu_{0j}, c_{0j}), \quad \pi(\sigma_j^2) \propto \frac{1}{\sigma_j^2}, \tag{2.11}$$

where μ_{0j} and c_{0j} represent the hyper-parameters of the distributions of μ_j, and $f(x) \propto g(x)$ means there exists a constant c such that $f(x) = c \cdot g(x)$. Even though there is not prior evidence to establish the value of the hyper-parameters, we selected $\mu_{0j} = 0$ and a very large value for c_{0j} to avoid bias. The distributions of μ_j and σ_j^2 will be iteratively updated based on the previous samples during the Gibbs sampling process.

2.4.1. Posterior Full Conditional Distributions

During the iterative Gibbs sampling process, the samples of different variables at each step are drawn based on their respective posterior distributions, conditional on current states of all other variables. Thus the implementation of Gibbs sampling requires the knowledge of the posterior full conditional distributions of all parameters of interests, including $\mathbf{q}$, $\{\mu_j\}_{j=1}^k$, $\{\sigma_j^2\}_{j=1}^k$, and $\mathbf{z}$.

Given the current state of $\mathbf{z}$, consider the set of pixels that correspond to the j-th component as $\mathcal{S}_j = \{i : z_i = j\}$ with cardinality $n_j = |\mathcal{S}_j|$. Define a vector $\mathbf{x}_j = [x_i]_{i \in \mathcal{S}_j} \in \mathcal{R}_+^{n_j \times 1}$, which contains x_i corresponding to all pixels labeled as $z_i = j$. We can update the conditional densities of $\boldsymbol{\theta}_j$ and $\mathbf{q}$ by using $\mathbf{z}$ and $\mathbf{x}_j$ as follows (See appendix 2.8.1 for the complete proof of the posterior distributions).

1. Posterior Distribution of $\mathbf{q}$

$$\pi(\mathbf{q}|\mathbf{z}) = \text{Dir}(\alpha_1 + n_1, \alpha_2 + n_2, ..., \alpha_k + n_k) \tag{2.12}$$

2. Posterior Distribution of σ_j^2

$$\pi(\sigma_j^2|\mu_j, \mathbf{x}_j) = \text{Inverse-Gamma}(a_j, b_j), \tag{2.13}$$

where $a_j = \frac{n_j}{2}$ and $b_j = \frac{1}{2}\sum_{i \in S_j}(x_i - \mu_j)^2$

3. Posterior Distribution of μ_j

$$f(\mu_j|\sigma_j^2, \mathbf{x}_j) = \mathcal{N}(m_j, v_j^2), \tag{2.14}$$

where $m_j = v_j^2\left(\frac{\sum_{i \in S_j} x_i}{\sigma_j^2} + \frac{\mu_0}{c_{0j}}\right)$ and $v_j^{-2} = \frac{n_j}{\sigma_j^2} + \frac{1}{c_{0j}}$

4. Posterior Distribution of z_i

$$\Pr\left(z_i = j|x_i, \{\boldsymbol{\theta}_j\}_{j=1}^k, \mathbf{q}\right) = \frac{f(x_i|\mu_j, \sigma_j^2, z_i = j)q_j}{\sum_{p=1}^k f(x_i|\mu_p, \sigma_p^2, z_i = p)q_p} \tag{2.15}$$

where $f(x_i|\mu_j, \sigma_j^2, z_i = j) = \mathcal{N}(\mu_j, \sigma_j^2)$.

2.4.2. Gibbs Sampling

In Gibbs sampling, we will iteratively draw samples from the posterior distributions derived in the previous subsections. The Gibbs sampling algorithm with Gaussian mixture model is summarized in Algorithm 2.

Algorithm 2 Gibbs sampling for Gaussian mixture model

Input: Data: $\mathbf{x}$; and hyper-parameters for prior distributions: $[\mu_{0j}, c_{0j}]_{j=1}^{k}$, $\boldsymbol{\alpha}$.
Initialization: Obtain $[\mu_j^{(0)}, \sigma_j^{(0)}]_{j=1}^{k}$, $\mathbf{z}^{(0)}$ by using results from K-means clustering.
for $it = 1, 2, \cdots, T$ **do**
  Formulate $\mathcal{S}_j^{(it)} = \{i : z_i^{(it-1)} = j\}$ and $\mathbf{x}_j^{(it)} = [x_i]_{i \in \mathcal{S}_i^{(it-1)}}$ by using $\mathbf{z}^{(it-1)}$
  Draw $\mathbf{q}^{(it)}$ from (2.12) using $\mathbf{z}^{(it-1)}$.
  Draw $(\sigma_j^2)^{(it)}$ from (2.13) using $\mu_j^{(it-1)}$ and $\mathbf{x}_j^{(it)}$, for $j = 1, \cdots, k$.
  Draw $\mu_j^{(it)}$ from (2.14) using $(\sigma_j^2)^{(it)}$ and $\mathbf{x}_j^{(it)}$, for $j = 1, \cdots, k$.
  Draw $z_i^{(it)}$ from (2.15) using $[\mu_j^{(it)}, (\sigma_j^2)^{(it)}]_{j=1}^{k}$, $\mathbf{q}^{(it)}$, and x_i, for $i \in \mathcal{S}_j^{(it)}$ and $j = 1, \cdots, k$.
end for
Output: $\mathbf{z}^{(it)}$, for $it = 1, \cdots, T$.

As the number of iterations grows large, the samples drawn through this process converge to their joint distributions. With such a process, the values of all model parameters are learned from the data without the need of a training process. The output of the Gibbs sampling algorithm is then used to evaluate the posterior probability $\Pr(z_i|\mathbf{x})$ and obtain an estimate on $\hat{z}_i$ as described in (2.8) and (2.9) in Algorithm 1.

2.5. Gibbs Sampling with t-Mixture Model

Gibbs sampling with t-mixture model is discussed in this section. With the t-mixture model, it is assumed that the likelihood function of x_i follows a t-distribution as follows,

$$f(x_i|\mu_j, \sigma_j^2, d_j, z_i = j) = t_{d_j}(\mu_j, \sigma_j^2), \tag{2.16}$$

where d_j represents the degree-of-freedom for the j-th component in the t-mixture model, and μ_j and σ_j^2 are the corresponding mean and variance, respectively. Instead of selecting a fixed value of d_j, we estimate the possible values of d_j from the data. The prior of d_j is assumed to be uniformly distributed with support $d_j \in \{d_{01}, d_{02}, \cdots, d_{0L}\}$ to produce both lighter and heavier tails. Since the direct application of the Gibbs sampler on the t-mixture model is challenging, we can rewrite

the model as [22]

$$f(x_i|\mu_j, \sigma_j^2, \lambda_i, z_i = j) = \mathcal{N}\left(\mu_j, \frac{\sigma_j^2}{\lambda_i}\right), \tag{2.17}$$

where λ_i is a parameter determined by the choice of d_j, for all $i \in \mathcal{S}_j$. The prior distributions of μ_j, σ_j^2, λ_i, and d_j are assumed to be

$$\pi(\mu_j) = \mathcal{N}(\mu_{0j}, c_{0j}), \text{ for } j = 1, \cdots, k, \tag{2.18}$$

$$\pi(\sigma_j^2) \propto \frac{1}{\sigma_j^2}, \text{ for } j = 1, \cdots, k, \tag{2.19}$$

$$\pi(\lambda_i|d_j) = \text{Gamma}\left(\frac{d_j}{2}, \frac{d_j}{2}\right), \text{ for } i \in \mathcal{S}_j,\ j = 1, \cdots, k, \tag{2.20}$$

$$\Pr(d_j = d_{0l}) = \frac{1}{L}, \text{ for } j = 1, \cdots, k, l = 1, \cdots, L \tag{2.21}$$

where μ_{0j} and c_{0j} represent the hyper-parameters of the distribution.

2.5.1. Posterior Full Conditional Distributions

With the model given before, the unknown model parameters are $\boldsymbol{\theta} = [\{\mu_j\}_{j=1}^k, \{\sigma_j^2\}_{j=1}^k, \{\lambda_i\}_{i=1}^N, \{d_j\}_{j=1}^k]$ and $\mathbf{q}$. We will use Gibbs sampling to draw samples of $\boldsymbol{\theta}$, $\mathbf{q}$, and $\mathbf{z}$ using their respective posterior full conditional distributions as follows (See appendix 2.8.2 for the complete proof of the posterior distributions).

1. Posterior Distribution of σ_j^2

$$f(\sigma_j^2|\mu_j, \boldsymbol{\lambda}_j, \mathbf{x}_j) = \text{Inverse-Gamma}(a_j, b_j), \tag{2.22}$$

   where $a_j = \frac{n_j}{2}$ and $b_j = \frac{1}{2}\sum_{i \in S_j} \lambda_i (x_i - \mu_j)^2$

2. Posterior Distribution of μ_j

$$f(\mu_j|\sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) = \mathcal{N}(m_j, v_j^2), \tag{2.23}$$

where $m_j = v_j^2 \left(\frac{\sum_{i \in S_j} \lambda_i x_i}{\sigma_j^2} + \frac{\mu_0}{c_{0j}} \right)$ and $v_j^{-2} = \frac{\sum_{i \in S_j} \lambda_i}{\sigma_j^2} + \frac{1}{c_{0j}}$

3. Posterior Distribution of d_j

$$\Pr(d_j = d_{0l} | \mu_j, \sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) = \frac{1}{C} \prod_{i \in \mathcal{S}_j} (\lambda_i)^{\frac{d_{0l}}{2} - 1} \exp\left(-\frac{d_{0l}}{2} \lambda_i \right), \tag{2.24}$$

where C is a normalization constant satisfying

$$C = \sum_{l=1}^{L} \prod_{i \in \mathcal{S}_j} (\lambda_i)^{\frac{d_{0l}}{2} - 1} \exp\left(-\frac{d_{0l}}{2} \lambda_i \right).$$

4. Posterior Distribution of λ_i

$$f(\lambda_i | \mu_j, \sigma_j^2, d_j, x_i) = \text{Gamma}\left(\frac{1 + d_j}{2}, \frac{1}{2\sigma_j^2}(x_i - \mu_j) + \frac{d_j}{2}^2 \right) \tag{2.25}$$

5. Posterior Distribution of z_i

$$\Pr(z_i = j | x_i, \boldsymbol{\theta}, \mathbf{q}) = \frac{f(x_i | \mu_j, \sigma_j^2, d_j, z_i = j) q_j}{\sum_{p=1}^{k} f(x_i | \mu_p, \sigma_p^2, d_j, z_i = p) q_p} \tag{2.26}$$

where $f(x_i | \mu_j, \sigma_j^2, d_p, z_i = j) = t_{d_j}(\mu_j, \sigma_j^2)$.

Considering that the posterior distribution of $\mathbf{q}$ depends only on $\mathbf{z}$, and it is independent of the specific mixture model, we can use the same distribution $f(\mathbf{q}|\mathbf{z})$ as in (2.12) for the t-mixture model.

2.5.2. Gibbs Sampling

The initialization of $\mathbf{z}^{(0)}$, $\mathbf{q}^{(0)}$, $\{\mu_j^{(0)}\}_{j=1}^k$, and $\{(\sigma_j^2)^{(0)}\}_{j=1}^k$ can be obtained by using the results from the K-means classification as described in Section 2.3.2. For the parameters $\{\lambda_i^{(0)}\}_{i=1}^N$ and $\{d_j^{(0)}\}$, we can initizalize them by draw i.i.d. random samples from their priori distributions in (2.20) and (2.21), respectively.

Algorithm 3 Gibbs sampling for t-mixture model

Input: Data: $\mathbf{x}$; and hyper-parameters for prior distributions: $[\mu_{0j}, c_{0j}]_{j=1}^{k}$, $\boldsymbol{\alpha}$.
Initialization:
Obtain $[\mu_j^{(0)}, \sigma_j^{(0)}]_{j=1}^{k}$, $\mathbf{z}^{(0)}$ by using results from K-means clustering.
Obtain $\{\lambda_i^{(0)}\}_{i=1}^{N}$ by randomly draw N i.i.d. samples with (2.20).
Obtain $\{d_j^{(0)}\}$ by randomly draw k i.i.d. samples with (2.21).
for $it = 1, 2, \cdots, T$ **do**
Formulate $\mathcal{S}_j^{(it)} = \{i : z_i^{(it-1)} = j\}$ and $\mathbf{x}_j^{(it)} = [x_i]_{i \in \mathcal{S}_i^{(it-1)}}$ by using $\mathbf{z}^{(it-1)}$
Draw $\mathbf{q}^{(it)}$ from (2.12) using $\mathbf{z}^{(it-1)}$.
Draw $(\sigma_j^2)^{(it)}$ from (2.22) using $\mu_j^{(it-1)}$, $\boldsymbol{\lambda}_j{}^{(it-1)}$, and $\mathbf{x}_j^{(it-1)}$, for $j = 1, \cdots, k$.
Draw $\mu_j^{(it)}$ from (2.23) using $(\sigma_j^2)^{(it)}$ $\boldsymbol{\lambda}_j{}^{(it-1)}$, and $\mathbf{x}_j^{(it-1)}$, for $j = 1, \cdots, k$.
Draw $d_j^{(it)}$ from (2.24) using $\boldsymbol{\lambda}_j{}^{(it-1)}$, for $j = 1, \cdots, k$.
Draw $\lambda_i^{(it)}$ from (2.25) using $[\mu_j^{(it)}, (\sigma_j^2)^{(it)}]$, $d_j^{(it)}$ and $\mathbf{x}_j^{(it-1)}$, for $i \in \mathcal{S}_j^{(it)}$, and $j = 1, \cdots, k$.
Draw $z_i^{(it)}$ from (2.26) using $[\mu_j^{(it)}, (\sigma_j^2)^{(it)}]_{j=1}^{k}$, $\{\lambda_i^{(it)}\}_{i=1}^{N}$, $\mathbf{q}^{(it)}$, and x_i, for $i \in \mathcal{S}_j^{(it)}$ and $j = 1, \cdots, k$.
end for
Output: $\mathbf{z}^{(it)}$, for $it = 1, \cdots, T$.

The Gibbs sampling algorithm with the t-mixture model is summarized in Algorithm 3. The output of the Gibbs sampling algorithm is then used to evaluate the emperical posterior probability $\Pr(z_i|\mathbf{x})$ and obtain an estimate on $\hat{z}_i$ as described in (2.8) and (2.9) in Algorithm 1.

2.6. Experimental Results

In this section, we describe the experimental results obtained from the implementation of the unsupervised algorithm described in sections 2.3, 2.4, and 2.5. Thirteen samples from mice breast cancer tumors were used to evaluate the performance of the proposed algorithm. Besides obtaining their corresponding pathology results, both fresh and FFPE THz images were collected from each of the samples. Some of the results were previously presented in [2] while for this chapter, we have selected three samples with either two or three types of tissue each. The hyper-parameters used in the analysis for all samples were: $\mu_{0j} = 0$ and $\sigma_{0j}^2 = 100$ for $j = 1, \cdots, k$. $\boldsymbol{\alpha} = [3, \cdots, 3]$, where $k = 2$ for samples 2 and 3, and $k = 3$ for samples 9B.

In order to evaluate the performance of the proposed algorithm, we compared the classification

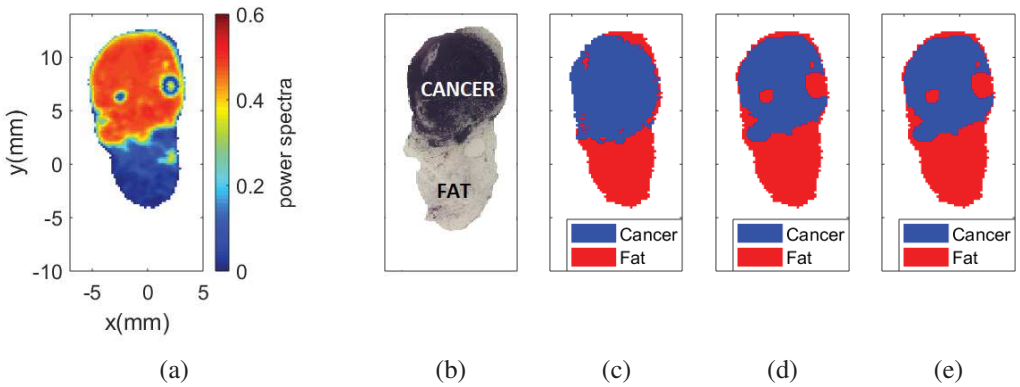


Figure 2.1: Sample 2 fresh. (a) THz image [2]. (b) Pathology image [2]. (c) Morphed Pathology. (d) Gaussian mixture model. (e) t-mixture model.

results with digitized pathology data, which were obtained through histopathology processing on the FFPE tissue block samples [2]. Since the pathology was performed on fixed tissue yet the THz imaging was obtained from fresh tissue, the shape of the pathology and THz images were slightly different and the pathology results presented a much higher resolution than the THz image. To make a pixel-by-pixel comparison possible, we morphed the shape of the pathology image into the contour of the THz image using mesh morphing [23]. For this purpose, the resolution of the pathology image was reduced and its contour was aligned to the THz image using their maximum correlation, similar to [2]. Once aligned, some key features in the contour of both images were selected to create a triangular shaped mesh and morph each triangle using homography estimation. It should be noted that during the morphing process, only the external contour and orientation of the THz image were used, and no feature inside the THz image was used to avoid artificial bias.

2.6.1. Samples with Two Types of Regions

The results obtained from freshly exercised tissues for samples 2 and 3 are shown in Figs. 2.1 and 2.2, respectively.

Similarly, Fig. 2.1a shows the THz reflected spectral power image for sample 2. Figs. 2.1b and 2.1c show the aligned pathology image and the morphed pathology image, respectively. From Fig. 2.1a we can observe that the cancer region produces a higher reflection in the THz image while

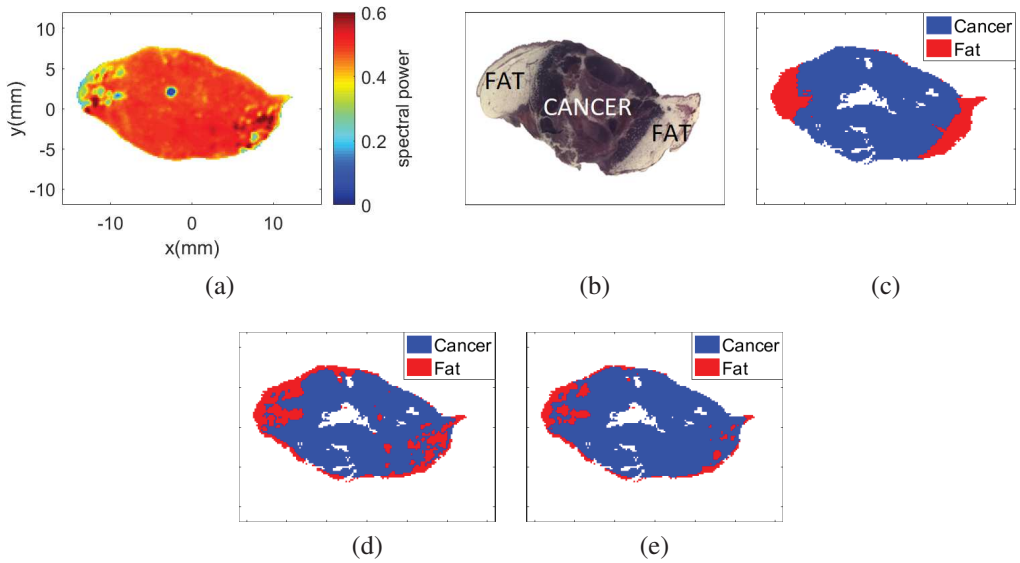


Figure 2.2: Sample 3 fresh. (a) THz image [2]. (b) Pathology image [2]. (c) Morphed Pathology. (d) Gaussian mixture model. (e) t-mixture model.

the fat regions generate a lower reflection, producing a strong visual agreement with the pathology image in Fig. 2.1c. The THz imaging classification results obtained by the proposed algorithm with the Gaussian mixture model are shown in Fig. 2.1d and the results obtained by using the t-mixture model are shown in Fig. 2.1e.

Fig. 2.2a shows the THz reflected spectral power image for sample 3. Fig. 2.2b shows the aligned pathology image, which represent the reference location of the different types of tissues within the tumor. Fig. 2.2c shows the morphed pathology image digitized to binary values based on the pathology results, with blue representing cancer and red representing fat. From Fig. 2.2a we can observe that the cancer region produces a higher reflection in the THz image while the fat regions generate a lower reflection, producing a good visual agreement with the pathology image in Fig. 2.2c. The THz imaging classification results obtained by the proposed algorithm with the Gaussian mixture model are shown in Fig. 2.2d and the results obtained by using the t-mixture model are shown in Fig. 2.2e.

To quantify the performance of the proposed algorithms, pixel-by-pixel comparisons were performed between the classification results (Figs. 2.2d and 2.2e for sample 3, Figs. 2.1d and 2.1e for

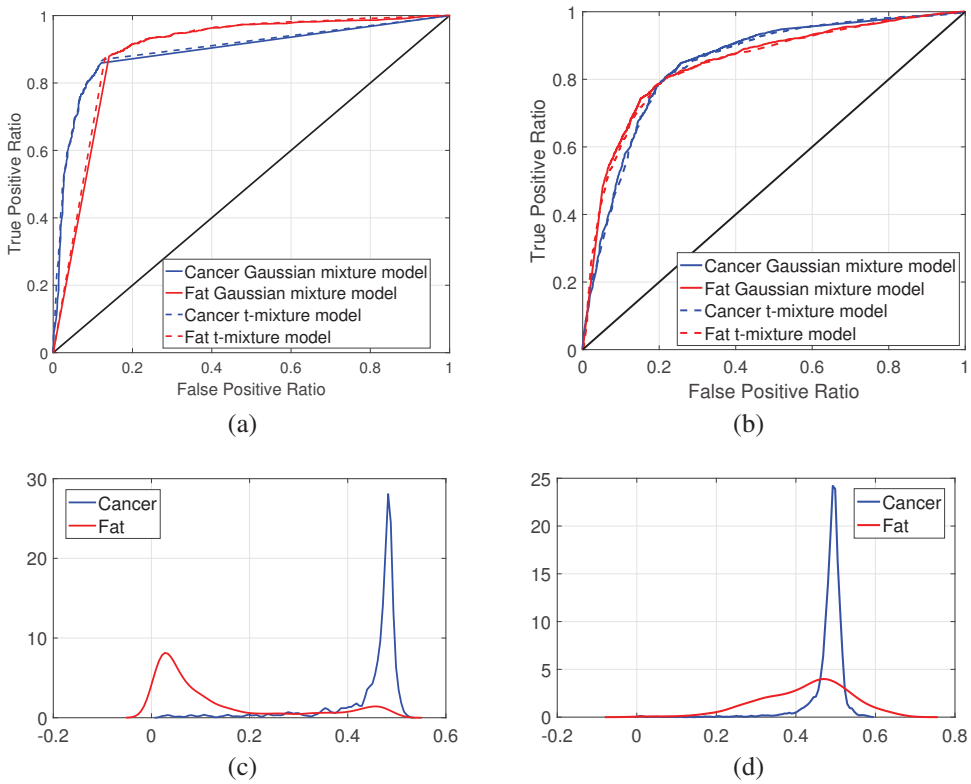


Figure 2.3: Samples 2 and 3 fresh. (a) ROC curves for sample 2 fresh. (b) ROC curves for sample 3 fresh. (c) Probability distribution for sample 2 fresh. (d) Probability distribution for sample 3 fresh.

sample 2) and their morphed pathology counterparts (Fig. 2.2c for sample 3, Fig. 2.1c for sample 2). The obtained receiver operating characteristic (ROC) curves are shown in Figs. 2.3a and 2.3b for samples 2 and 3, respectively. For cancer detection in sample 3, the algorithm can achieve a true positive ratio of 80% at a false positive ratio of 20%; for cancer detection in sample 2 the algorithm can achieve a true positive ratio of 83% at a false positive ratio of 10%.

To better understand the performance difference between samples 2 and 3, we plotted the empirical mixture distributions in Figs. 2.3c and 2.3d with the Gaussian mixture model. For sample 2, the distribution peaks of the two components in the mixture models are apart from each other; on the other hand, for sample 3, the distribution peaks are very close to each other. This is due to excess fluid accumulating between the tissue and polystyrene imaging plate. Since water (and by extent blood and PBS) has very high absorption in the THz range, the distributed fluid across the

Table 2.1: Area under the ROC curve for fresh samples 2 and 3.

Type of tissue: Model	Sample 2	Sample 3
Cancer: Gaussian mixture model	0.8930	0.8492
Fat: Gaussian mixture model	0.8930	0.8492
Cancer: t-mixture model	0.8989	0.8462
Fat: t-mixture model	0.8989	0.8462

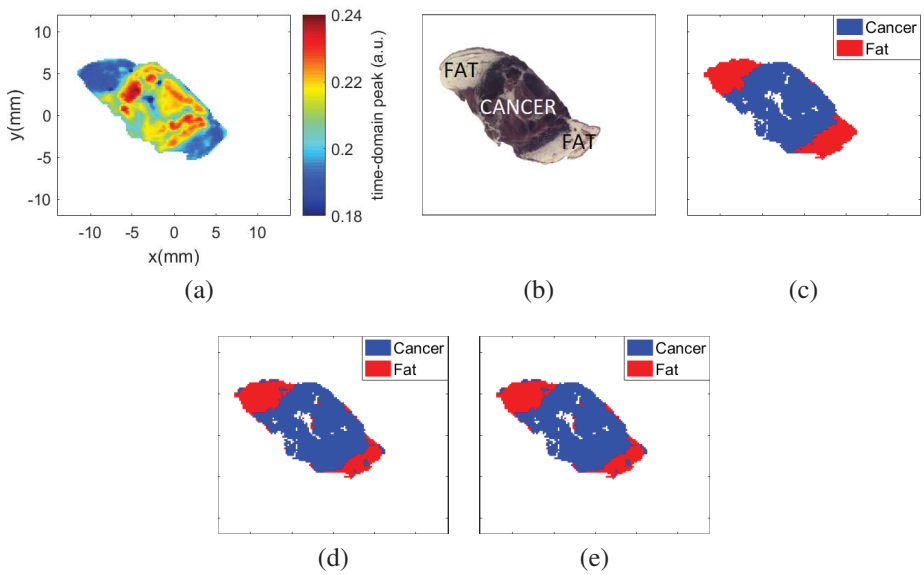
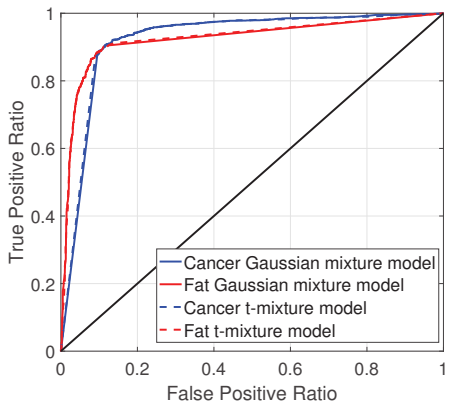


Figure 2.4: Sample 3 block. (a) THz image. (b) Pathology image [2]. (c) Morphed Pathology. (d) Gaussian mixture model. (e) t-mixture model.

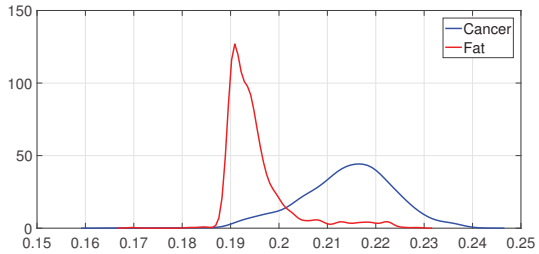
fat regions of sample 3 creates a higher reflection similar to cancer. Therefore it is more difficult to separate cancer from fat in sample 3, while the tissue regions in sample 2 were discrete with more thorough clearing of excess fluids.

The normalized area underneath the ROC curves can be used to measure the classification quality. An area of 100% corresponding to the perfect ROC curve of 100% true positive at 0% false positive. The normalized areas of the ROC curves in Figs. 2.3a and 2.3b are given in Table 2.1. All areas are greater than 84%. For sample 2, the t-mixture model outperforms the Gaussian mixture model, and the normalized area can reach 89.89% with the t-mixture model. For sample 3, the Gaussian-mixture model and t-mixture model have similar performances.

Next we study the classification performance obtained by analyzing the THz imaging of the



(a)



(b)

Figure 2.5: Sample 3 FFPE. (a) ROC curves. (b) Probability distribution.

FFPE tissue of sample 3, which is shown in Fig. 2.4a. Fig. 2.4b presents the aligned pathology image, and Fig. 2.4c shows the morphed digitized pathology image that was used to evaluate the performance of the algorithms. While there is some change in the surface between the THz image of the fresh tissue and the pathology image due to histopathology processing, the THz image of the FFPE tissue is taken at the same surface as pathology and therefore shows very close agreement. Figs. 2.4d and 2.4e represent the classification results with the Gaussian and t-mixture models, respectively.

The ROC curves and empirical distributions of the classification results of FFPE sample 3 are shown in Figs. 2.5a and 2.5b, respectively. The normalized ROC areas are 92.36% and 92.53% for Gaussian and t-mixture models, respectively. It should be noted that for tissue with two components, the normalized ROC areas are the same for both cancer and fat. As can be seen from Fig. 2.5b, the two components in the mixture model of the FFPE sample are better separated than their counterparts on the fresh sample shown in Fig. 2.3d, thus the classification results of the FFPE

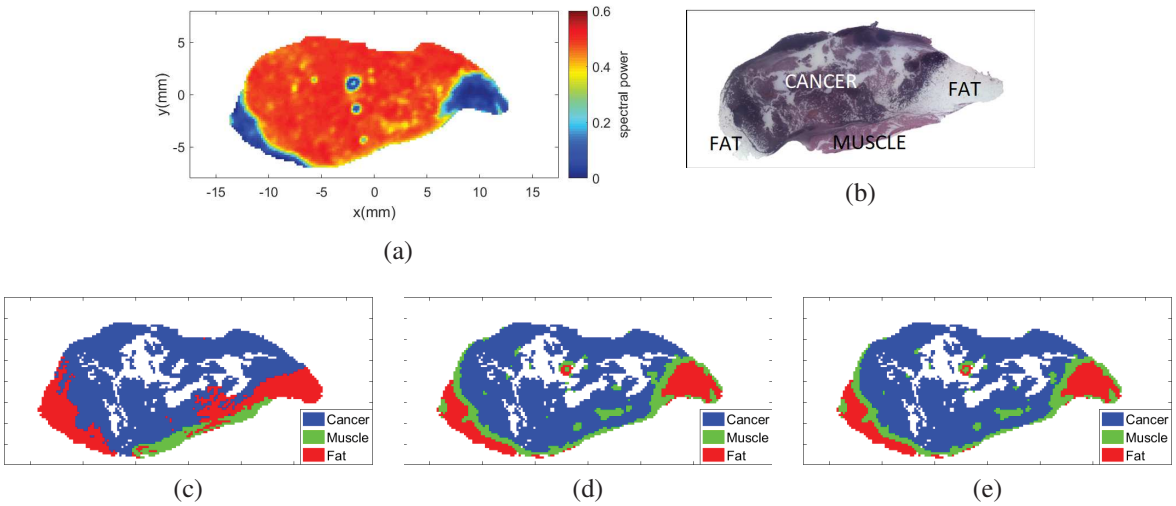


Figure 2.6: Sample 9B fresh. (a) THz image [24]. (b) Pathology image [24]. (c) Morphed Pathology [24]. (d) Gaussian mixture model [24]. (e) t-mixture model.

sample is better than the fresh sample. In addition, the t-mixture model slightly outperforms the Gaussian mixture model.

2.6.2. Samples with Three Types of Regions

Next we study the performance of the algorithm with respect to sample 9B, which has cancer, fat, and muscle. Fig. 2.6a shows the THz reflected spectral power image for sample 9B. Figs. 2.6b and 2.6c present the aligned pathology image and the corresponding digitized morphed pathology image. From Fig. 2.6a, we can observe that while the regions of cancer and fat are readily apparent in a qualitative sense, the region of muscle is indistinct from cancer. Although there may be slightly lower reflection where we assume the muscle is based on pathology (refer to Fig. 2.6c), this is not conclusive. Figs. 2.6d and 2.6e represent the classification results using Gaussian and t-mixture models, respectively. In this case, when the model is attempting to consider three different tissue regions, there is a tendency to pick up the transition between high and low reflections as a third tissue region which makes distinguishing between regions with close reflection properties like muscle and cancer a challenge.

The ROC curves and empirical distributions of the classification results of fresh sample 9B

Table 2.2: Area under the ROC curve for sample 9B.

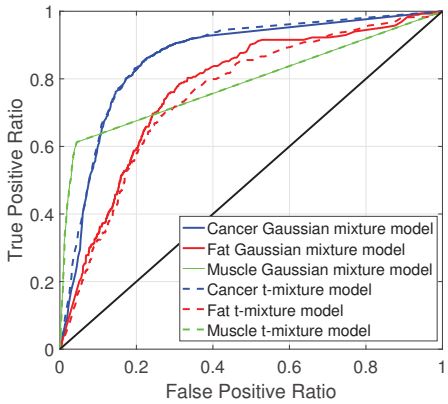
Type of tissue	Gaussian mixture model	t-mixture model
Cancer	0.8630	0.8698
Muscle	0.7729	0.7534
Fat	0.7885	0.7885

are shown in Figs. 2.7a and 2.7b, respectively. The normalized ROC areas are given in Table 2.2. As can be seen from the distributions in Fig. 2.7b, the distributions of muscle and cancer are close to each other, and the distribution of fat has two peaks, with one peak very close to that from cancer. As a result, the ROC curves in Fig. 2.7a show that the detection performance of muscle is challenging, and the performance with respect to fat is compromised due to the model picking up the transition between cancer and fat regions. However, the algorithm can still obtain reasonably good performance with respect to cancer with a normalized ROC area of 86.98% with the t-mixture model.

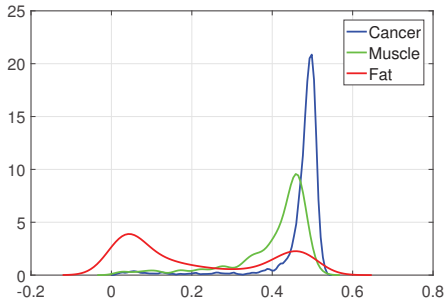
Upon comparing the distributions of the mixture model seen in Figs. 2.3d, 2.3c, 2.5b, and 2.7b, we can see that the distributions of different tissue regions are well separated, while they partially overlap otherwise. The more challenging case is seen for sample 9B, where the THz reflections from the muscle and cancer regions have insignificant difference. Therefore, we will investigate models with higher degree-of-freedom in our future work to achieve better classifications of regions with close reflection properties.

2.7. Conclusion

Unsupervised Bayesian learning algorithms have been developed for cancer detection in THz imaging of freshly exercised murine tissues. Under a Bayesian framework, the algorithms were developed by using Gaussian mixture or t-mixture models, where each component in the mixture model was used to represent one type of tissue. The model parameters and posterior distributions were iteratively learned from the data by using MCMC with Gibbs sampling. Experimental results indicated that the t-mixture model slightly outperforms the Gaussian mixture model, and both can



(a)



(b)

Figure 2.7: Sample 9B fresh. (a) ROC curves. (b) Probability distribution.

identify different tissue types by using THz imaging with satisfactory accuracy. The primary challenges in the model-based classification have to do with regions where excess fluid may alter the reflections from the tissue, as with the fat in sample 3, or when regions have very similar or overlapping reflections, as with the cancer and muscle in sample 9B.

In our future work, one of the challenges that we plan to address is to further improve the identification accuracy when there are 3 or more types of tissues, in particular when it includes muscle or fibrous tissues. Given that the probability distribution for muscle and cancer are similar, it is necessary to add more parameters or features to the model in order to achieve better distinction between the two. Another important direction is to improve feature extraction such that we can use multiple dimensional features instead of just intensity to summarize the properties of different tissue types. Some promising signal features of pulsed THz imaging that can lead to better classification have been identified via principal component analysis in [8], so a similar investigation will be conducted in the future.

In summary, the current results of our algorithm are promising and they represent the first steps to achieve an accurate approach for breast cancer detection through THz imaging.

2.8. Appendix

2.8.1. Posterior updates for Gaussian mixture model

This appendix presents the mathematical proofs of (2.12)-(2.15), which represent the posterior distributions of the variables involved in the Gibbs sampling for the Gaussian mixture model.

Posterior Distribution of $\mathbf{q}$

In Bayesian statistics, the parameter $\mathbf{q}$ of the latent variable $\mathbf{z}$ is usually assumed to follow Dirichlet distribution with parameter $\boldsymbol{\alpha}$ as $\pi(\mathbf{q}) = \text{Dir}(\boldsymbol{\alpha})$. Based on the assumption of Dirichlet prior and the Bayes' rule, the posterior distribution of $\mathbf{q}$ can be written as

$$f(\mathbf{q}|\mathbf{z}) = \frac{f(\mathbf{z}|\mathbf{q})\pi(\mathbf{q})}{f(\mathbf{z})} \propto \prod_{j=1}^{k} q_j^{n_j} \prod_{j=1}^{k} q_j^{\alpha_j - 1} = \prod_{j=1}^{k} q_j^{(\alpha_j + n_j) - 1}$$

where n_j is the number of pixels labeled in the j-th category by $\mathbf{z}$. Therefore, the posterior distribution of $\mathbf{q}$ still follows Dirichlet distribution as

$$\pi(\mathbf{q}|\mathbf{z}) = \text{Dir}(\alpha_1 + n_1, \alpha_2 + n_2, ..., \alpha_k + n_k)$$

Posterior Distribution of μ_j

Based on Bayes' rule, we have

$$f(\mu_j|\sigma_j^2, \mathbf{x}_j) = \frac{f(\mathbf{x}_j|\mu_j, \sigma_j^2)\pi(\mu_j)\pi(\sigma_j^2)}{\pi(\sigma_j^2, \mathbf{x}_j)} \propto f(\mathbf{x}_j|\mu_j, \sigma_j^2)\pi(\mu_j)\pi(\sigma_j^2) \tag{2.27}$$

Substituting (2.10) and (2.11) into (2.27) yields

$$f(\mu_j|\sigma_j^2, \mathbf{x}_j) \propto \left(\frac{1}{\sqrt{2\pi}}\right)^{n_j} \sigma_j^{-n_j} \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i\in S_j}(x_i-\mu_j)^2\right\} \frac{1}{\sqrt{2\pi c_{0j}}} \exp\left\{-\frac{(\mu_j-\mu_{0j})^2}{2c_{0j}}\right\} \frac{1}{\sigma_j^2} \tag{2.28}$$

If we only consider the terms that include μ_j, then

$$f(\mu_j|\sigma_j^2, \mathbf{x}_j) \propto \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i\in S_j}(x_i-\mu_j)^2 - \frac{(\mu_j-\mu_{0j})^2}{2c_{0j}}\right\},$$

which can be equivalently written as

$$f(\mu_j|\sigma_j^2, \mathbf{x}_j) \propto \exp\left\{-\frac{1}{2}\left(\frac{\sigma_j^2+n_jc_{0j}}{\sigma_j^2c_{0j}}\right)\left(\mu_j - \left(\frac{c_{0j}\sum_{i\in S_j}x_i+\sigma_j^2\mu_{0j}}{\sigma_j^2+n_jc_{0j}}\right)\right)^2\right\}$$

Therefore, the posterior of μ_j can be represented as:

$$f(\mu_j|\sigma_j^2, \mathbf{x}_j) = \mathcal{N}(m_j, v_j^2),$$

where $m_j = v_j^2\left(\frac{\sum_{i\in S_j}x_i}{\sigma_j^2} + \frac{\mu_0}{c_{0j}}\right)$ and $v_j^{-2} = \frac{n_j}{\sigma_j^2} + \frac{1}{c_{0j}}$.

Posterior Distribution σ_j^2

Similar to (2.27), based on Bayes' rule, the posterior distribution of σ_j^2 can be calculated as

$$f(\sigma_j^2|\mu_j, \mathbf{x}_j) \propto f(\mathbf{x}_j|\mu_j, \sigma_j^2)\pi(\mu_j)\pi(\sigma_j^2),$$

which is proportional to the right-hand-side (RHS) of (2.28). If we only consider the terms that include σ_j^2 in (2.28), then

$$f(\sigma_j^2|\mu_j, \mathbf{x}_j) \propto \left(\sigma_j^2\right)^{-\frac{n_j}{2}-1} \exp\left\{-\frac{1}{\sigma_j^2}\frac{\sum_{i\in S_j}(x_i-\mu_j)^2}{2}\right\}$$

Therefore, the posterior distribution of σ_j^2 follows the inverse Gamma distribution as

$$\pi(\sigma_j^2|\mu_j, \mathbf{x}_j) = \text{Inverse-Gamma}(a_j, b_j),$$

where $a_j = \frac{n_j}{2}$ and $b_j = \frac{1}{2}\sum_{i\in S_j}(x_i-\mu_j)^2$.

Posterior Distribution of z_i

Consider the i-th pixel with $i \in \mathcal{S}_j$. Given the values of the parameters $\boldsymbol{\theta}_j = [\mu_j, \sigma_j^2]$, $\mathbf{q}$, and the pixel intensity x_i, we can calculate the posterior probability mass function (PMF) of z_i as

$$\Pr\left(z_i = j|x_i, \{\boldsymbol{\theta}_j\}_{j=1}^k, \mathbf{q}\right) = \frac{f(x_i|\mu_j, \sigma_j^2, z_i = j)q_j}{\sum_{p=1}^k f(x_i|\mu_p, \sigma_p^2, z_i = p)q_p}$$

Therefore, the posterior distribution of z_i follows multinomial distribution with the parameters being

$$q_j' = \frac{f(x_i|\mu_j, \sigma_j^2, z_i = j)q_j}{\sum_{p=1}^k f(x_i|\mu_p, \sigma_p^2, z_i = p)q_p}, \quad j = 1, \cdots, k,$$

where $f(x_i|\mu_j, \sigma_j^2, z_i = j) = \mathcal{N}(\mu_j, \sigma_j^2)$.

2.8.2. Posterior updates for t-mixture model

This appendix presents the mathematical proofs of (2.22)-(2.26), which represent the posterior distributions of the variables involved in the Gibbs sampling for the Gaussian mixture model.

Distribution of μ_j

Define $\boldsymbol{\lambda}_j = [\lambda_i]_{i \in \mathcal{S}_j}$. Based on Bayes' rule, the posterior distribution of μ_j can be written as

$$f(\mu_j|\sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) \propto f(\mathbf{x}_j|\mu_j, \sigma_j^2, \boldsymbol{\lambda}_j)\pi(\mu_j)\pi(\sigma_j^2)\pi(\boldsymbol{\lambda}_j)$$

Based on (2.17)-(2.20), the posterior distribution can be written as

$$f(\mu_j|\sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) \propto \prod_{i \in \mathcal{S}_j} \left\{ \sqrt{\frac{\lambda_i}{2\pi\sigma_j^2}} \exp\left[-\frac{\lambda_i}{2\sigma_j^2}(x_i - \mu_j)^2 \right] \pi(\lambda_i) \right\} \frac{1}{\sqrt{2\pi c_{0j}}} \exp\left\{ -\frac{(\mu_j - \mu_{0j})^2}{2c_{0j}} \right\} \frac{1}{\sigma_j^2} \tag{2.29}$$

Considering only the terms that include μ_j yields

$$f(\mu_j|\sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) \propto \exp\left\{ -\frac{1}{2\sigma_j^2} \sum_{i \in S_j} (x_i - \mu_j)^2 \lambda_i - \frac{(\mu_j - \mu_{0j})^2}{2c_{0j}} \right\},$$

which can be alternatively written as

$$f(\mu_j|\sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) \propto \exp\left\{ -\frac{1}{2} \left(\frac{\sigma_j^2 + c_{0j} \sum_{i \in S_j} \lambda_i}{\sigma_j^2 c_{0j}} \right) \left(\mu_j - \left(\frac{c_{0j} \sum_{i \in S_j} x_i \lambda_i + \sigma_j^2 \mu_{0j}}{\sigma_j^2 + c_{0j} \sum_{i \in S_j} \lambda_i} \right) \right)^2 \right\}.$$

Therefore, the posterior distribution of μ_j can be represented as

$$f(\mu_j|\sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) = \mathcal{N}(m_j, v_j^2),$$

where $m_j = v_j^2 \left(\frac{\sum_{i \in S_j} \lambda_i x_i}{\sigma_j^2} + \frac{\mu_0}{c_{0j}} \right)$ and $v_j^{-2} = \frac{\sum_{i \in S_j} \lambda_i}{\sigma_j^2} + \frac{1}{c_{0j}}$.

Posterior Distribution of σ_j^2

Based on Bayes' rule, the posterior distribution of σ_j^2 can be written as

$$f(\sigma_j^2|\mu_j, \boldsymbol{\lambda}_j, \mathbf{x}_j) \propto f(\mathbf{x}_j|\mu_j, \sigma_j^2, \boldsymbol{\lambda}_j)\pi(\mu_j)\pi(\sigma_j^2)\pi(\boldsymbol{\lambda}_j)$$

where the RHS is the same as that of (2.29). Removing the terms that are irrelevant to σ_j^2, we have

$$f(\sigma_j^2|\mu_j, \boldsymbol{\lambda}_j, \mathbf{x}_j) \propto (\sigma_j^2)^{-\frac{n_j}{2}-1} \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i\in S_j}(x_i-\mu_j)^2\lambda_i\right\}$$

Therefore, the posterior distribution can be described as,

$$f(\sigma_j^2|\mu_j, \boldsymbol{\lambda}_j, \mathbf{x}_j) = \text{Inverse-Gamma}(a_j, b_j),$$

where $a_j = \frac{n_j}{2}$ and $b_j = \frac{1}{2}\sum_{i\in S_j}\lambda_i(x_i-\mu_j)^2$.

Posterior Distribution of λ_i

Assume $z_i = j$, that is, $i \in \mathcal{S}_j$. The posterior distribution of λ_i can be expressed as

$$f(\lambda_i|\mu_j, \sigma_j^2, d_j, x_i) \propto f(x_i|\mu_j, \sigma_j^2, \lambda_i, d_j)\pi(\mu_j)\pi(\sigma_j^2)\pi(\lambda_i|d_j)\pi(d_j).$$

Considering only the terms that include λ_i yields

$$f(\lambda_i|\mu_j, \sigma_j^2, d_j, x_i) \propto (\lambda_i)^{\frac{1}{2}} \exp\left\{-\frac{(x_i-\mu_j)^2\lambda_i}{2\sigma_j^2}\right\}(\lambda_i)^{\frac{d_j}{2}-1}\exp\left(-\frac{d_j}{2}\lambda_i\right),$$

which can be alternatively expressed as

$$f(\lambda_i|\mu_j, \sigma_j^2, d_j, x_i) \propto (\lambda_i)^{\frac{d_j+1}{2}-1} \exp\left\{-\left(\frac{(x_i-\mu_j)^2}{2\sigma_j^2}+\frac{d_j}{2}\right)\lambda_i\right\}$$

Therefore, the posterior distribution of λ_i follows the Gamma distribution as

$$f(\lambda_i|\mu_j, \sigma_j^2, d_j, x_i) = \text{Gamma}\left(\frac{1+d_j}{2}, \frac{1}{2\sigma_j^2}(x_i - \mu_j) + \frac{d_j}{2}^2\right)$$

Posterior Distribution of d_j

The posterior PMF of d_j can be written as

$$\Pr(d_j = d_{0l}|\mu_j, \sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) \propto f(\mathbf{x}_j|\mu_j, \sigma_j^2, \boldsymbol{\lambda}_j)\pi(\mu_j)\pi(\sigma_j^2)\pi(\boldsymbol{\lambda}_j|d_j = d_{0l})\pi(d_j = d_{0l})$$

After removing the terms irrelevant to d_j, we have

$$\Pr(d_j = d_{0l}|\mu_j, \sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) \propto \prod_{i \in \mathcal{S}_j} (\lambda_i)^{\frac{d_{0l}}{2}-1} \exp\left(-\frac{d_{0l}}{2}\lambda_i\right).$$

Therefore, the posterior PMF of d_j can be evaluated as

$$\Pr(d_j = d_{0l}|\mu_j, \sigma_j^2, \boldsymbol{\lambda}_j, \mathbf{x}_j) = \frac{1}{C}\prod_{i \in \mathcal{S}_j} (\lambda_i)^{\frac{d_{0l}}{2}-1} \exp\left(-\frac{d_{0l}}{2}\lambda_i\right),$$

where C is a normalization constant satisfying

$$C = \sum_{l=1}^{L} \prod_{i \in \mathcal{S}_j} (\lambda_i)^{\frac{d_{0l}}{2}-1} \exp\left(-\frac{d_{0l}}{2}\lambda_i\right).$$

Posterior distribution of z_i

Given the values of the parameters $\boldsymbol{\theta} = [\{\mu_j\}_{j=1}^k, \{\sigma_j^2\}_{j=1}^k, \{\lambda_i\}_{i=1}^N, \{d_j\}_{j=1}^k]$, $\mathbf{q}$, and the pixel intensity x_i, we can calculate the posterior distribution of z_i as

$$\Pr(z_i = j|x_i, \boldsymbol{\theta}, \mathbf{q}) = \frac{f(x_i|\mu_j, \sigma_j^2, d_j, z_i = j)q_j}{\sum_{p=1}^k f(x_i|\mu_p, \sigma_p^2, d_j, z_i = p)q_p}$$

Therefore, the posterior distribution of z_i follows multinomial distribution with the parameters being

$$q'_j = \frac{f(x_i|\mu_j, \sigma_j^2, d_j, z_i = j)q_j}{\sum_{p=1}^{k} f(x_i|\mu_p, \sigma_p^2, d_j, z_i = p)q_p}, \quad j = 1, \cdots, k.$$

where $f(x_i|\mu_j, \sigma_j^2, d_j, z_i = j) = t_{d_j}(\mu_j, \sigma_j^2)$.

References

[1] Breastcancer.org. (2018) U.S. Breast Cancer Statistics. [Online]. Available: http://www.breastcancer.org/symptoms/understand_bc/statistics

[2] T. Bowman, T. Chavez, K. Khan, J. Wu, A. Chakraborty, N. Rajaram, K. Bailey, and M. El-Shenawee, "Pulsed terahertz imaging of breast cancer in freshly excised murine tumors," *Journal of Biomedical Optics*, vol. 23, no. 2, p. 026004, 2018. doi: 10.1117/1.JBO.23.2.026004

[3] T. C. Bowman, M. El-Shenawee, and L. K. Campbell, "Terahertz imaging of excised breast tumor tissue on paraffin sections," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 5, pp. 2088–2097, May 2015. doi: 10.1109/TAP.2015.2406893

[4] T. Bowman, M. El-Shenawee, and L. K. Campbell, "Terahertz transmission vs reflection imaging and model-based characterization for excised breast carcinomas," *Biomed. Opt. Express*, vol. 7, no. 9, pp. 3756–3783, Sept. 2016. doi: 10.1364/BOE.7.003756. [Online]. Available: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-7-9-3756

[5] T. Bowman, Y. Wu, J. Gauch, L. K. Campbell, and M. El-Shenawee, "Terahertz imaging of three-dimensional dehydrated breast cancer tumors," *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 38, no. 6, pp. 766–786, Mar. 2017. doi: 10.1007/s10762-017-0377-y

[6] T. Bowman, A. Walter, O. Shenderova, N. Nunn, G. McGuire, and M. El-Shenawee, "A phantom study of terahertz spectroscopy and imaging of micro- and nano-diamonds and nano-onions as contrast agents for breast cancer," *Biomedical Physics & Engineering Express*, vol. 3, no. 5, p. 055001, 2017. [Online]. Available: http://stacks.iop.org/2057-1976/3/i=5/a=055001

[7] A. J. Fitzgerald, V. P. Wallace, M. Jimenez-Linan, L. Bobrow, R. J. Pye, A. D. Purushotham, and D. D. Arnone, "Terahertz pulsed imaging of human breast tumors," *Radiology*, vol. 239, no. 2, pp. 533–540, 2006. doi: 10.1148/radiol.2392041315 PMID: 16543586. [Online]. Available: https://doi.org/10.1148/radiol.2392041315

[8] A. J. Fitzgerald, V. P. Wallace, S. E. Pinder, A. D. Purushotham, P. O'Kelly, and P. C. Ashworth, "Classification of terahertz-pulsed imaging data from excised breast tissue," *Journal of Biomedical Optics*, vol. 17, no. 1, 2012. doi: 10.1117/1.JBO.17.1.016005. [Online]. Available: http://dx.doi.org/10.1117/1.JBO.17.1.016005

[9] L. H. Eadie, C. B. Reid, A. J. Fitzgerald, and V. P. Wallace, "Optimizing multi-dimensional terahertz imaging analysis for colon cancer diagnosis," *Expert Systems with Applications*, vol. 40, no. 6, pp. 2043 – 2050, 2013. doi: https://doi.org/10.1016/j.eswa.2012.10.019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412011335

[10] S. Roostaie, H. Kashanian, H. Ghaffari, N. B. Ajami, and H. Alidoost, "Gastric cancer diagnosis using terahertz imaging," *Majlesi Journal of Multimedia Processing*, vol. 4, no. 4, 2016. [Online]. Available: http://www.mp.majlesi.info/index/index.php/mjmm/article/view/183

[11] Z. Fengyu, L. Ming, Y. Lei, and Y. Xianfeng, "Image segmentation algorithm of Gaussian mixture model based on map/reduce," in *2017 Chinese Automation Congress (CAC)*, Oct. 2017, pp. 1520–1525. doi: 10.1109/CAC.2017.8243008

[12] J. Tian and K.-K. Ma, "A MCMC approach for Bayesian super-resolution image reconstruction," in *IEEE International Conference on Image Processing 2005*, vol. 1, Sept. 2005. ISSN 1522-4880 pp. I–45–8. doi: 10.1109/ICIP.2005.1529683

[13] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, Mar. 2002. doi: 10.1109/34.990138

[14] N. Dobigeon, S. Moussaoui, M. Coulon, J. Y. Tourneret, and A. O. Hero, "Subspace-based Bayesian blind source separation for hyperspectral imagery," in *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec. 2009, pp. 372–375. doi: 10.1109/CAMSAP.2009.5413255

[15] Y. Altmann, S. McLaughlin, and A. Hero, "Robust linear spectral unmixing using anomaly detection," *IEEE Transactions on Computational Imaging*, vol. 1, no. 2, pp. 74–85, June 2015. doi: 10.1109/TCI.2015.2455411

[16] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov chain Monte Carlo in practice*, 1st ed. London: Chapman & Hall, 1996. ISBN 9780412055515;0412055511;

[17] X. Shen, C. R. Dietlein, E. Grossman, Z. Popovic, and F. G. Meyer, "Detection and segmentation of concealed objects in terahertz images," *IEEE Transactions on Image Processing*, vol. 17, no. 12, pp. 2465–2475, Dec. 2008. doi: 10.1109/TIP.2008.2006662

[18] J. Diebolt and C. P. Robert, "Estimation of finite mixture distributions through Bayesian sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 56, no. 2, pp. 363–375, 1994. [Online]. Available: http://0-www.jstor.org.library.uark.edu/stable/2345907

[19] A. E. Gelfand, *Gibbs Sampling*. John Wiley & Sons, Inc., 2004. ISBN 9780471667193. [Online]. Available: http://dx.doi.org/10.1002/0471667196.ess0302.pub2

[20] B. S. Peter, S. Yngvesson, P. Siqueira, P. Kelly, A. Khan, S. Glick, and A. Karellas, "Development and testing of a single frequency terahertz imaging system for breast cancer detection," *IEEE Transactions on Terahertz Science and Technology*, vol. 3, no. 4, pp. 374–386, July 2013. doi: 10.1109/TTHZ.2013.2241429

[21] M. Stephens, "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 62, no. 4, pp. 795–809, 2000. [Online]. Available: http://0-www.jstor.org.library.uark.edu/stable/2680622

[22] C. Fernández and M. F. J. Steel, "Bayesian regression analysis with scale mixtures of Normals," *Econometric Theory*, vol. 16, no. 1, pp. 80–101, 2000. [Online]. Available: http://www.jstor.org/stable/3533160

[23] G. Wolberg, "Recent advances in image morphing," in *Computer Graphics International, 1996. Proceedings*, June 1996, pp. 64–71. doi: 10.1109/CGI.1996.511788

[24] T. Chavez, T. Bowman, J. Wu, M. El-Shenawee, and 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*, July 2019. ISSN 1522-3965 pp. 525–526. doi: 10.1109/APUSNCURSINRSM.2019.8888653

Chapter 3

Breast Cancer Detection with Low-dimension Ordered Orthogonal Projection in Terahertz Imaging

Tanny Chavez, Nagma Vohra, Jingxian Wu, Keith Bailey, and Magda El-Shenawee

3.1. Abstract

This chapter proposes a new dimension reduction algorithm based on low-dimension ordered orthogonal projection (LOOP), 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 dimension spectrum vector covering several THz frequencies, where each frequency represents a different dimension in the vector. The proposed algorithm projects the high-dimension spectrum vector of each pixel within the THz image into a low-dimension subspace that contains the majority of the unique features embedded in the image. The low-dimension 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-dimension 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. Additionally, we explore the introduction of the spatial correlation among neighboring pixels in THz images through a Markov random field (MRF) approach. Experiment results demonstrate that the proposed method achieves apparent performance improvement in human breast cancer tissue over existing approaches such as one- dimension Markov chain Monte Carlo. The results confirm that the dimension reduction algorithm presented in this chapter 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.

3.2. 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 have 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,

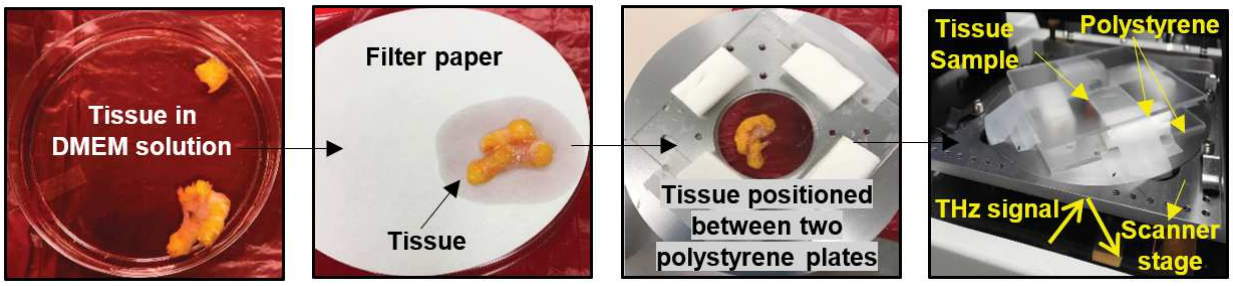


Figure 3.1: Sample preparation process. (a) The tissue immersed in DMEM solution, (b) removal of excess water in tissue using filter paper, (c) the tissue positioned in a sandwich between two polystyrene plates, and (d) positioning the tissue sandwich on scanner stage for imaging.

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 (SVM) [11–13, 15], probabilistic neural networks (PNN) [12, 13], and deep neural networks (DNN) [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 dimension spectrum vector covering several THz frequencies, where each frequency represents a different dimension in the vector. The high-dimension 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, 8], the high-dimension THz waveform per pixel is summarized into a scalar, such as the peak of the reflected time-domain signal or the energy over a certain frequency band. The one-dimension (1D) feature extractions used in [5, 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 (LPP) [12].

In this chapter we propose a new low-dimension ordered orthogonal projection (LOOP) algorithm, which is used to extract low-dimension features embedded in the high-dimension THz waveform. The low-dimension feature extraction is achieved by projecting the original THz signal into a low-dimension subspace containing the majority of the salient information necessary for classification. The low-dimension 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-dimension 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 dimension 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 chapter 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 chapter focuses on human breast tumor samples.

The low-dimension feature vector is analyzed and modeled by using a multivariate Gaussian mixture model (GMM) [18], with each component in GMM corresponding to one possible tissue type within the sample. The prevalence of different tissue types within a sample are modeled by using the weight or prior probability for each component in 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 GMM with the low-dimension 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 vs. healthy tissue) [12], this chapter 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.

In addition, considering that the pixels are collected by scanning the tumor with steps of $200\mu m$, it is natural to assume that neighboring pixels have a higher probability of belonging to the same region. Motivated by this fact, we propose a spatial image segmentation algorithm that exploits the spatial correlation among pixels. The spatial correlation is modeled by applying Markov random field (MRF) on GMM, and the EM algorithm is then applied to the statistical models to classify the different regions in the THz image.

The rest of the chapter is organized as follows. Section 3.3 presents the experiment setup and data collection process. Section 3.4 introduces the problem formulation and notations used in the chapter. Details of the LOOP algorithm are explained in Section 3.5. Section 3.6 defines the GMM and the unsupervised learning algorithms based on the low-dimension vector obtained by LOOP. Section 3.8.3 describes the introduction of the spatial information per pixel. Section 3.8 shows the experimental results, and section 3.9 concludes the chapter.

3.3. Experiment Setup

The experimental set-up was established in Dr. El-Shenawee's Terahertz Imaging and Spectroscopy Lab at the University of Arkansas. The raw experimental data was provided by Ms. Nagma Vohra, PhD candidate in Dr. El-Shenawee's group.

The tissue samples handled in this work follow the Environmental Health and Safety protocol of the University of Arkansas. The experimental work done in this chapter makes use of human breast cancer tissues # ND10898, ND15526, and ND15588 obtained from the National Disease Research Interchange within 24h of excision. These samples were obtained via left breast masectomy from a 59-year-old patient diagnosed with stage III/III infiltrating dual carcinoma (IDC), radical masectomy from a 90-year-old patient with stage III/III IDC, and masectomy from a 63-year-old patient with stage II/III IDC, respectively. On receiving the tissue in the Terahertz lab, it was removed from the Dulbecco's Modified Eagle Medium (DMEM) solution, see Fig. 3.1a. After removing excess water using filter paper (Fig. 3.1b), 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. 3.1c. This arrangement of the tissue was then mounted on the scanner stage for the reflection imaging procedure as shown in Fig. 3.1d.

The reflection measurements were taken by using a TPS Spectra 3000 pulsed THz imaging and spectroscopy system (from TeraView Ltd., UK). The diagram of the system is shown in Fig. 3.2a. The system uses a Ti:Sapphire laser that produces a 800 nm pulse to excite the THz emitter and THz receiver. Upon excitation, the THz emitter generates a time domain THz pulse as shown in Fig. 3.2b. The Fourier transform of the pulse, as shown in Fig. 3.2c, demonstrates a power spectra of pulse ranging from 0.1 THz 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 towards the THz receiver [8]. In the reflection mode measurements, both the THz emitter and 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

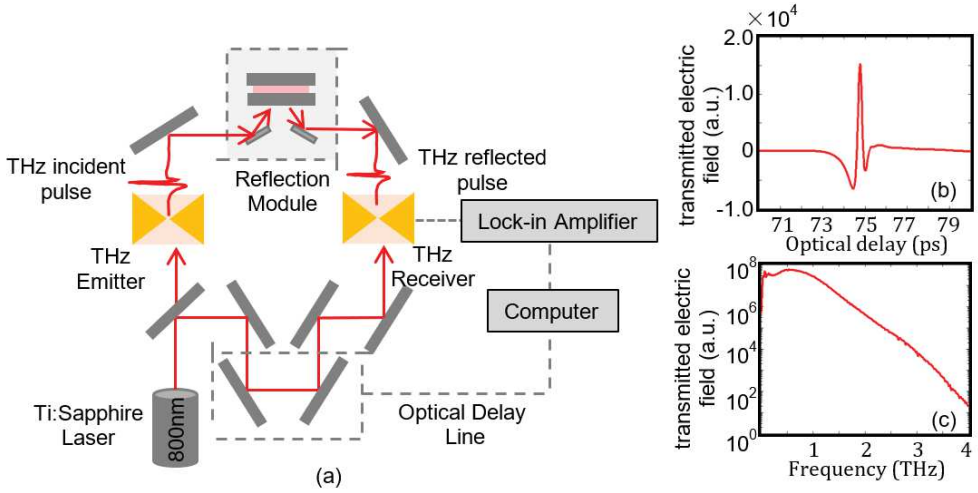


Figure 3.2: THz system description. (a) THz system diagram for reflection imaging, (b) incident time domain THz pulse, and (c) frequency spectrum of terahertz pulse in (b).

a stepper motor. The total time span of the imaging process was ∼30-40 minutes. 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 x-y 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 (OADDL) for the pathology process. The histopathology process involves fixing the tissue in formalin and embedding it in paraffin blocks. Further, from the formalin fixed paraffin embedded (FFPE) tissue blocks, two $\sim$ 3-4 μm thick slices were cut, stained with hematoxylin and eosin (H&E), 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 3.8.

3.4. Problem formulation

The problem formulation and notations are described in this section. Let the tensor $\mathcal{W} \in \mathcal{R}^{N_1 \times N_2 \times T}$ represent the THz image of size $N_1 \times N_2$. Each pixel $\mathcal{W}_{n_1,n_2,*}$ 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 axis, respectively.

For simplicity, the tensor $\boldsymbol{\mathcal{W}}$ is unfolded into a matrix $\boldsymbol{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_N] \in \mathcal{R}^{T \times N}$ 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 i-th pixel is $\mathbf{y}_i = \mathcal{F}(\mathbf{w}_i)$, where $\mathcal{F}(\cdot)$ is the FFT operator. Since $\mathbf{w}_i$ is real, the FFT of $\mathbf{w}_i$ is even-symmetric. Thus the size of $\mathbf{y}_i$ 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 or 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 spectrum. For analysis of the complex spectrum, we have $\mathbf{Y} = \mathbf{D}$; for analysis of the amplitude spectrum, $\mathbf{Y}$ is obtained by replacing all elements in $\mathbf{D}$ with their respective amplitudes.

We will perform cancer detection by utilizing the frequency-domain THz matrix $\mathbf{Y}$, 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 i-th pixel is represented by the frequency-domain vector $\mathbf{y}_i$, which has a relatively large dimension of $F = 106$. The frequency domain vector $\mathbf{y}_i$ 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 $\mathbf{y}_i$ means 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-dimension feature extraction before classification. With low-dimension feature extraction, the high dimension vector $\mathbf{y}_i$ can be mapped to a low dimension 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.

3.5. Low-Dimension Ordered Orthogonal Projection

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

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

ing 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 $\mathbf{Y}$ 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

$$\mathbf{b}_1 = \frac{\bar{\mathbf{y}}}{\|\bar{\mathbf{y}}\|} \tag{3.1}$$

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

The subsequent orthonormal basis vectors are calculated in a sequential manner. Assume the first l orthonormal basis vectors have been identified, and they are represented as $\mathcal{B}_l = [\mathbf{b}_1, \cdots, \mathbf{b}_l]$. The $(l+1)$-th basis vector will be calculated by using the pixel that is least represented by $\mathbf{B}_l$. 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-degree angle means the vector can be fully represented by the subspace.

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

$$P_{\mathcal{B}_l}(\mathbf{y}_i) = \sum_{j=1}^{l}\langle \boldsymbol{y}_i, \boldsymbol{b}_j\rangle \boldsymbol{b}_j \tag{3.2}$$

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

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

$$\cos(\theta_{i,l}) = \frac{\left\langle \boldsymbol{y}_i, P_{\mathcal{B}_l}(\mathbf{y}_i)\right\rangle}{\|\boldsymbol{y}_i\| \cdot \|P_{\mathcal{B}_l}(\mathbf{y}_i)\|}, \tag{3.3}$$

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

$$\boldsymbol{u}_{l+1} = \underset{\mathbf{y}_i \in \mathcal{Y}_l}{\text{argmin}} \, |\cos(\theta_{i,l})|. \tag{3.4}$$

where $\mathcal{Y}_l$ contains all the $\mathbf{y}_i$ 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, $\mathbf{b}_{l+1}$, can then be calculated by following the GS procedure as

$$\mathbf{v}_{l+1} = \boldsymbol{u}_{l+1} - P_{\mathcal{B}_l}(\boldsymbol{u}_{l+1}), \tag{3.5}$$

$$\mathbf{b}_{l+1} = \frac{\mathbf{v}_{l+1}}{\|\mathbf{v}_{l+1}\|}. \tag{3.6}$$

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 dimension representation of the THz image. Define $\mathbf{B}_L = [\mathbf{b}_1, \cdots, \mathbf{b}_L] \in \mathcal{C}^{F \times L}$, then the low dimension representation of $\mathbf{y}_i$ can be expressed as

$$\mathbf{y}_i = \mathbf{B}_L \times \mathbf{z}_i, \text{ for } i = 1, \cdots, N. \tag{3.7}$$

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

$$\mathbf{Y} = \mathbf{B}_L \times \mathbf{Z}, \tag{3.8}$$

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

3.6. Unsupervised Learning with Gaussian Mixture Model

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

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

$$\mathbf{X} = [\Re(\mathbf{Z}^T), \Im(\mathbf{Z}^T)]^T \in \mathcal{R}^{2L\times N} \tag{3.9}$$

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

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

$$f\left(\mathbf{x}_i|[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, q_k]_{k=1}^K\right) = \sum_{k=1}^{K} q_k g(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.10}$$

where K is the number of categories in the sample tissue, q_k is the prior probability of a pixel in the k-th 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 (pdf) 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 i-th pixel, for $i = 1, \cdots, N$. That is, $\zeta_i = k$ indicates that the i-th pixel belongs to the k-th category. It is assumed that the latent variable ζ_i follows a multinomial distribution with prior probability $\pi(\zeta_i = k) = q_k$, for $k = 1, \cdots, K$.

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

$$\hat{\zeta}_i = \underset{k\in\{1,\ldots,K\}}{\text{argmax}} \ \Pr(\zeta_i = k|\mathbf{X}) \tag{3.11}$$

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 chapter to obtain the classification results.

3.6.1. Markov Chain Monte Carlo

The posterior probability $\Pr(\zeta_i = k|\mathbf{X})$ 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 k-th category, and $n_k^{(0)} = \left|\mathcal{S}_k^{(0)}\right|$ is the cardinality of $\mathcal{S}_k^{(0)}$. The initial values of the variables can then be calculated as

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

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

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

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

$$\pi(q_k) = \text{Dir}(\alpha_k),$$

$$\pi(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_{0k}, \boldsymbol{\Sigma}_{0k}),$$

$$\pi(\boldsymbol{\Sigma}_k) = \text{InvWish}_p(\boldsymbol{\Psi}, \nu),$$

where Dir and InvWish represent the Direchlet and Inverse-Wishart distributions, respectively, α_k, $\boldsymbol{\mu}_{0k}$, $\boldsymbol{\Sigma}_{0k}$, $\boldsymbol{\Psi}$, and ν are the hyper-parameters 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]:

- Posterior distribution of $\boldsymbol{q}$

$$q_k \sim \text{Dir}(\alpha_k + n_k) \tag{3.12}$$

  where n_k is the number of pixels classified into the k-th category in the previous iteration.

- Posterior distribution of $\boldsymbol{\Sigma}_k$

$$\boldsymbol{\Sigma}_k \sim \text{InvWish}_p(\boldsymbol{S} + \boldsymbol{\Psi}, n_k + \nu) \tag{3.13}$$

  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 k-th category in the previous iteration.

- Posterior distribution of $\boldsymbol{\mu}_k$

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \tag{3.14}$$

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

- Posterior distribution of ζ_i

$$\Pr\left(\zeta_i = k \middle| \boldsymbol{x}_i, [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, q_k]_{k=1}^K\right) = \frac{q_k g(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \zeta_i = k)}{\sum_{p=1}^K q_p g(\boldsymbol{x}_i|\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p, \zeta_i = p)} \tag{3.15}$$

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,

$$\Pr(\zeta_i = k|\mathbf{X}) = \lim_{M\to\infty} \frac{1}{M} \sum_{it=1}^{M} \mathcal{I}(\zeta_i^{(it)} = k), \tag{3.16}$$

where $\mathcal{I}(a) = 1$ if $a =$ TRUE and 0 otherwise. MAP detection can then be applied with (3.16) to perform classification. It should be noted that, before applying the MCMC algorithm, the data vector $\mathbf{X}$ 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 chapter, the vectors $\mathbf{X}$ are scaled by a factor of 15 before applying the amplitude MCMC algorithm to fresh samples.

3.6.2. 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, $\ell(\boldsymbol{\theta}) = \log p(\mathbf{X}|\boldsymbol{\theta})$, as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} q_k g(\boldsymbol{x}_i|\zeta_i = k, \boldsymbol{\theta})\right) \tag{3.17}$$

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].

E-step

In the E-step of the m-th iteration, the algorithm first calculates the posterior probability of ζ_i by using (3.15), and the result is denoted as

$$\gamma_{ik}^{(m)} = \Pr\left(\zeta_i = k \middle| \boldsymbol{x}_i, \boldsymbol{\theta}^{(m)}\right) \tag{3.18}$$

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

M-step

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

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{N} \mathbb{E}_{\zeta_i|\boldsymbol{\theta}^{(m)}} \left[\log p\left(\boldsymbol{x}_i, \zeta_i \middle| \boldsymbol{\theta}\right)\right] \tag{3.19}$$

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

Calculating the conditional expectation in (3.19) yields

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = \sum_{k=1}^{K} \eta_k^{(m)} \left[\log q_k - \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_k|\right] - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik}^{(m)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k), \tag{3.20}$$

where

$$\eta_k^{(m)} = \sum_{i=1}^{N} \gamma_{ik}^{(m)}. \tag{3.21}$$

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

- Estimator of q_k

$$q_k^{(m+1)} = \frac{\eta_k^{(m)}}{N} \tag{3.22}$$

- Estimator of $\boldsymbol{\mu}_k$

$$\boldsymbol{\mu}_k^{(m+1)} = \frac{1}{\eta_k^{(m)}} \sum_{i=1}^{N} \gamma_{ik}^{(m)} \boldsymbol{x}_i \tag{3.23}$$

- Estimator of $\boldsymbol{\Sigma}_k$

$$\boldsymbol{\Sigma}_k^{(m+1)} = \frac{1}{\eta_k} \sum_{i=1} \gamma_{ik}^{(m)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(m+1)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(m+1)}\right)^T \tag{3.24}$$

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

3.7. Spatial Prior with Markov Random Field

Unlike the previous sections that assumed statistical independence among pixels in the spatial domain, this section proposes a novel spatial image segmentation approach that considers the spatial correlation among pixels. Specifically, it is assumed that pixels within a certain neighborhood, that is, a cluster of pixels that are close to each other, are correlated with each other by following certain prior distributions.

Let $\mathbf{Y} = [y_1, \ldots, y_N]$ denote the classification labels for the N pixels in the THz image, where $y_i \in \{1, 2, \ldots, K\}$, and K represent the number of regions (e.g. cancer, fat, collagen, etc.). The spatial correlation among the pixels can be represented by a Gibbs prior to the labels as [27]

$$P(\mathbf{Y}) = \frac{1}{Z} \exp\left(-\sum_{c \in \mathcal{C}} V_c(\mathbf{Y})\right), \tag{3.25}$$

where Z is a normalization constant, $\mathcal{C}$ corresponds to the clique within the defined neighborhood, $V_c(y_i, y_j) = \beta(1 - I_{y_i,y_j})$, and $I_{y_i,y_j} = 1$ if $y_i = y_j$ or 0 otherwise. For this purpose, we consider neighborhoods of sizes 4, 8, and 24 directly surrounding each pixel of interest. Finally, this new objective function is solved by using EM and GMM, similarly to the previous sections.

3.8. 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

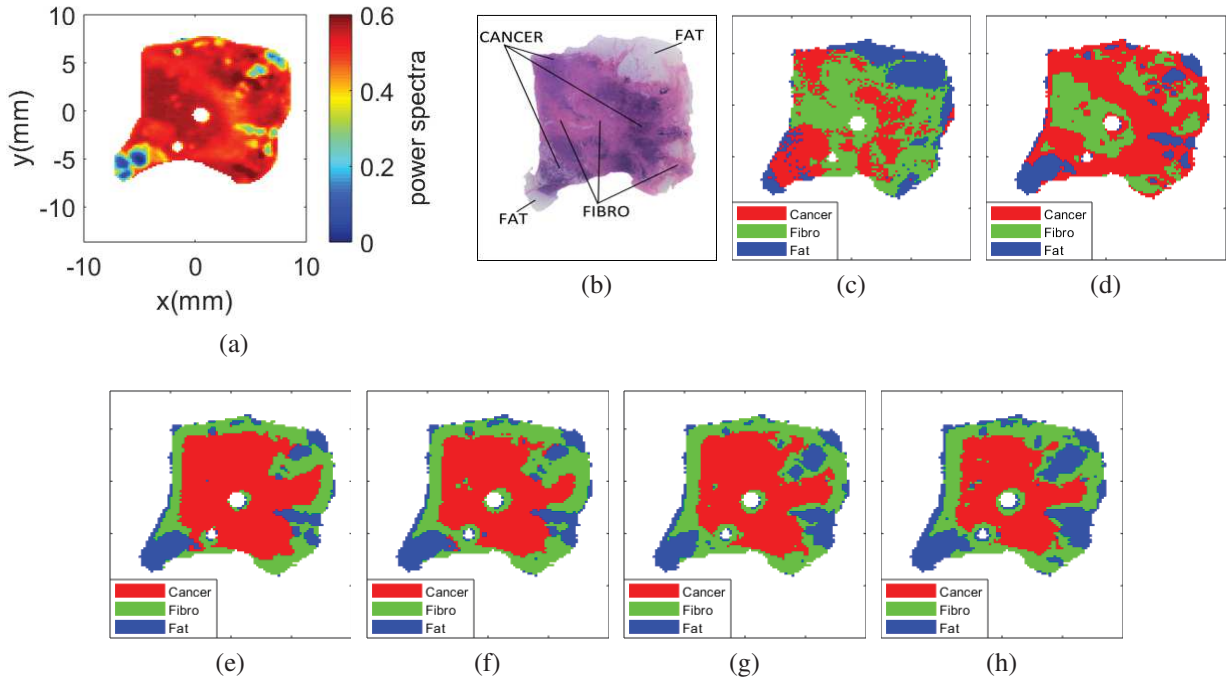


Figure 3.3: Sample ND10898 fresh. (a) THz image. (b) Pathology image. (c) Morphed Pathology. (d) 1D MCMC model. (e) 2D amplitude MCMC model. (f) 2D amplitude EM model. (g) 4D complex MCMC model. (h) 4D complex EM model.

source codes used for this analysis are available in [28].

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

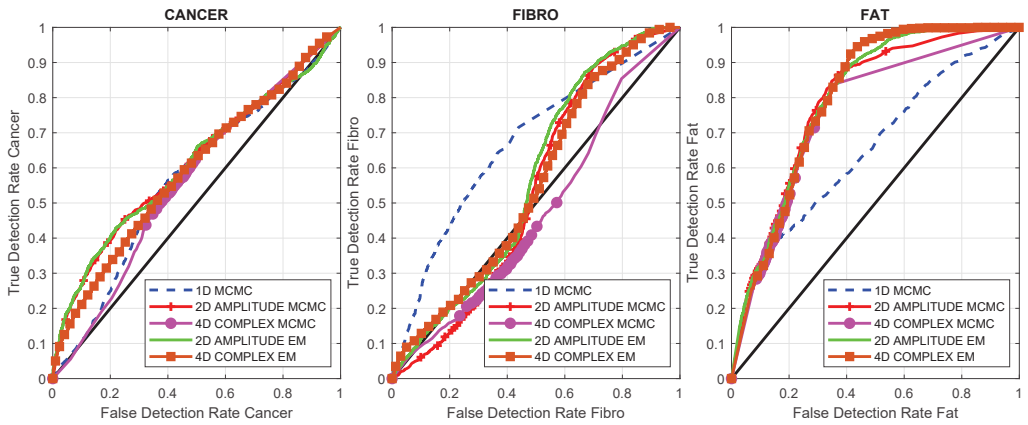


Figure 3.4: ROC curves for sample ND10898 fresh.

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 dimension 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 dimension 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 1D MCMC algorithm that summarizes each pixel into a 1D scalar[5, 8], and the PCA algorithm[29]. It is important to mention that the 1D 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.

3.8.1. 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 $15 \times 15mm$, $8.7 \times 13mm$, and $8 \times 15.3mm$, respectively. We receive fresh tissue of thickness ranging from 3 to 4mm. As reported in[7], the

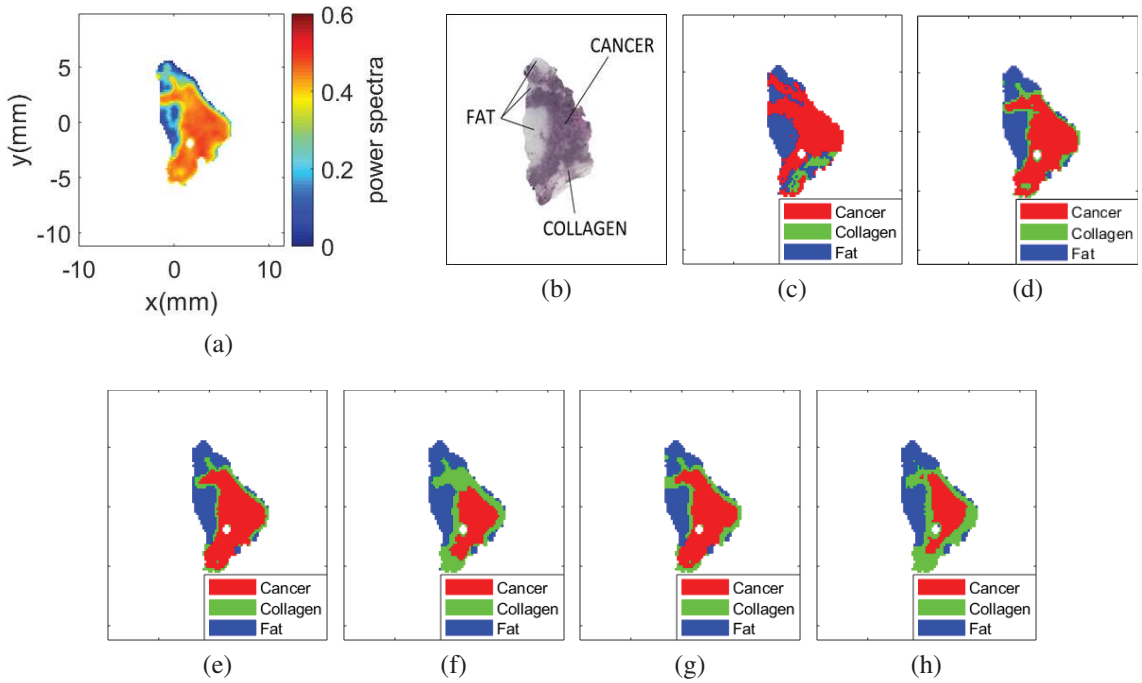


Figure 3.5: Sample ND15526 fresh. (a) THz image. (b) Pathology image. (c) Morphed Pathology. (d) 1D MCMC model. (e) 2D amplitude MCMC model. (f) 2D amplitude EM model. (g) 3D complex MCMC model. (h) 6D complex EM model.

tissue have high absorption coefficient ranging from $\sim$100 to 700cm^{-1} in the frequency range 0.1 to 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$276μm in cancer[30], therefore the reflected signal from tissue of less than $\sim$2mm thickness could be adversely affected by the multiple reflection. These samples contain three regions: cancer, collagen or fibro, and fat.

Fig. 3.3a 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. 3.3b represents the histopathology results obtained by analyzing the FFPE tissue sample fixed in paraffin, which corresponds to the gold standard within cancer detection. Fig. 3.3c 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. 3.3 represent air bubbles (artifact from the data collection process) that were removed before further processing to avoid data contamination.

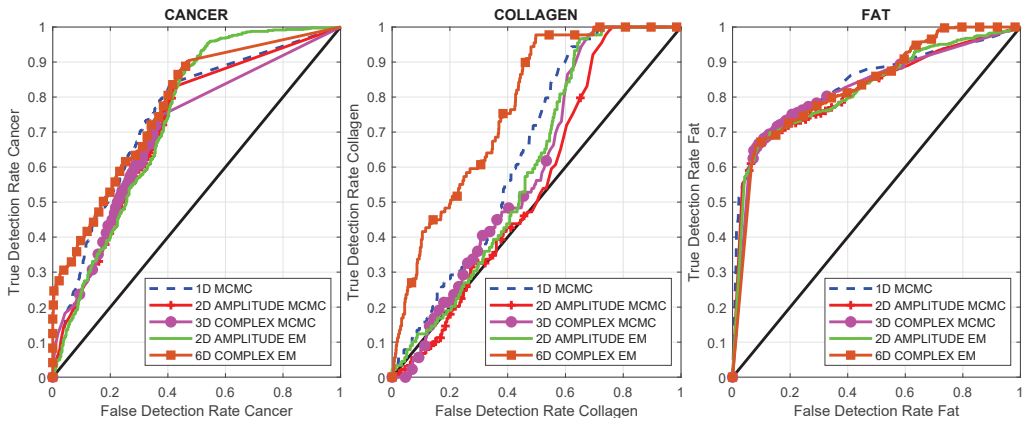


Figure 3.6: ROC curves for sample ND15526 fresh.

The classification results of the THz image obtained by using the 1D MCMC approach[5, 8], 4D MCMC with amplitude spectrum, 2D EM with amplitude spectrum, 4D MCMC with complex spectrum, and 4D EM with complex spectrum are presented in Figs. 3.3d-3.3h, respectively. The 2D and 4D 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 1D 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 2D and 4D models presented in Figs. 3.3e-3.3h when compared to the morphed pathology results in Fig. 3.3c.

To quantify the performance of each model, the corresponding ROC curves of the classification results of sample ND10898 fresh are presented in Fig. 3.4. The ROC curves are obtained by performing pixel-by-pixel comparisons between the detection results and the morphed pathlogy results. The areas underneath the ROC curves are listed in Table 4.1. All results obtained with the proposed LOOP algorithm perform significantly better than the 1D MCMC approach[8] for both cancer and fat, while the detection of fibro is better in 1D MCMC. The results from 2D feature vectors achieve larger cancer ROC areas ($\sim 60\%$) than those from the 1D approach ($\sim 50\%$). 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, 2D amplitude EM achieves the best performance for cancer and fat detection.

Similarly, Fig. 3.5a represents the THz image collected from sample ND15526 fresh. Figs.

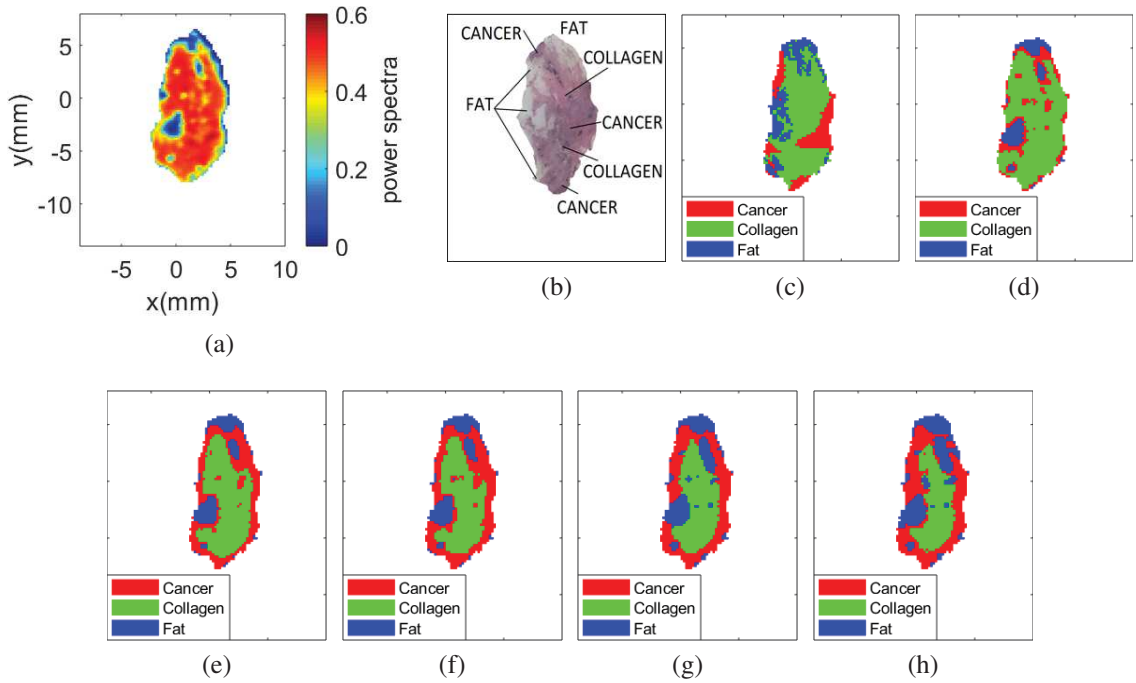


Figure 3.7: Sample ND15588 fresh. (a) THz image. (b) Pathology image. (c) Morphed Pathology. (d) 1D MCMC model. (e) 2D amplitude MCMC model. (f) 2D amplitude EM model. (g) 3D complex MCMC model. (h) 4D complex EM model.

3.5b and 3.5c correspond to the original and morphed histopathology results. Figs. 3.5d-3.5h show the classification results for 1D MCMC, 2D amplitude MCMC, 2D amplitude EM, 3D complex MCMC, and 6D complex EM, respectively. Visually, 1D MCMC, 2D amplitude MCMC, and 3D complex MCMC present similar classification areas with good cancer correlation, but with poor collagen detection. On the contrary, 2D amplitude EM and 6D complex EM present a better collagen detection at the cost of large cancer regions misclassification.

Fig. 3.6 presents the ROC curves for sample ND15526 fresh and their areas under the ROC curves are presented in Table 4.1. 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 1D MCMC being slightly better. Overall the best classification results are obtained by the 6D complex EM approach. This method achieved areas under the ROC of 77% or above for all the regions presented in this sample.

Fig. 3.7a shows the THz image collected from sample ND15588 while it was still fresh. Fig.

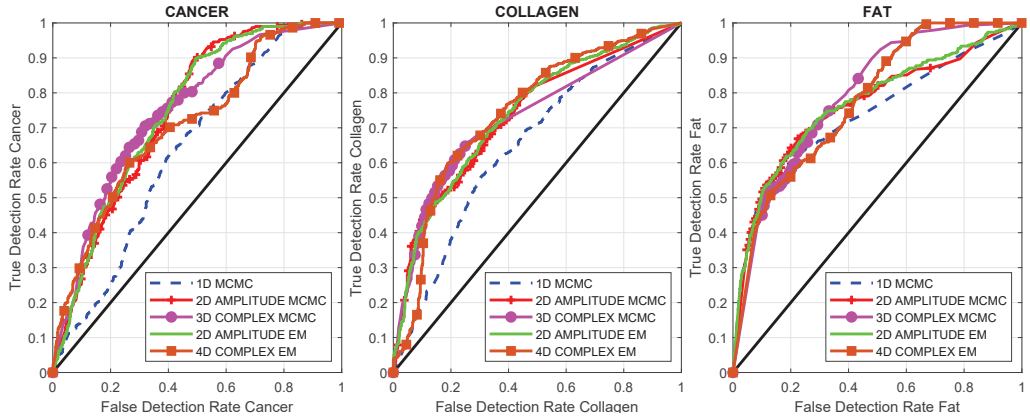


Figure 3.8: ROC curves for sample ND15588 fresh.

3.7b represents the histopathology results obtained by analyzing the corresponding FFPE tissue sample fixed in paraffin. Fig. 3.7c shows the morphed pathology mask obtained by employing the morphing algorithm[5]. The classification results obtained by using the 1D MCMC approach, 4D MCMC with amplitude spectrum, 2D EM with amplitude spectrum, 4D MCMC with complex spectrum, and 4D EM with complex spectrum are presented in Figs. 3.7d-3.7h, respectively. For the 1D MCMC approach, large portions of the cancer regions are misclassified as collagen. For the 2D and 4D 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-dimension 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 Figs. 3.7e-3.7h 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. 3.8. The areas underneath the ROC curves are listed in Table 4.1. All results obtained with the proposed LOOP algorithm perform considerably better than the 1D MCMC approach[8]. The results from 2D, 3D and 4D feature vectors achieve larger ROC areas ($\sim 70\%$) than those from the 1D approach ($\sim 60\%$). Thus we can conclude that increasing the dimension of the feature vector by just one dimension over the 1D approach can achieve apparent performance improvement. In terms of areas underneath the ROC curves, 2D amplitude EM achieves the best performance for

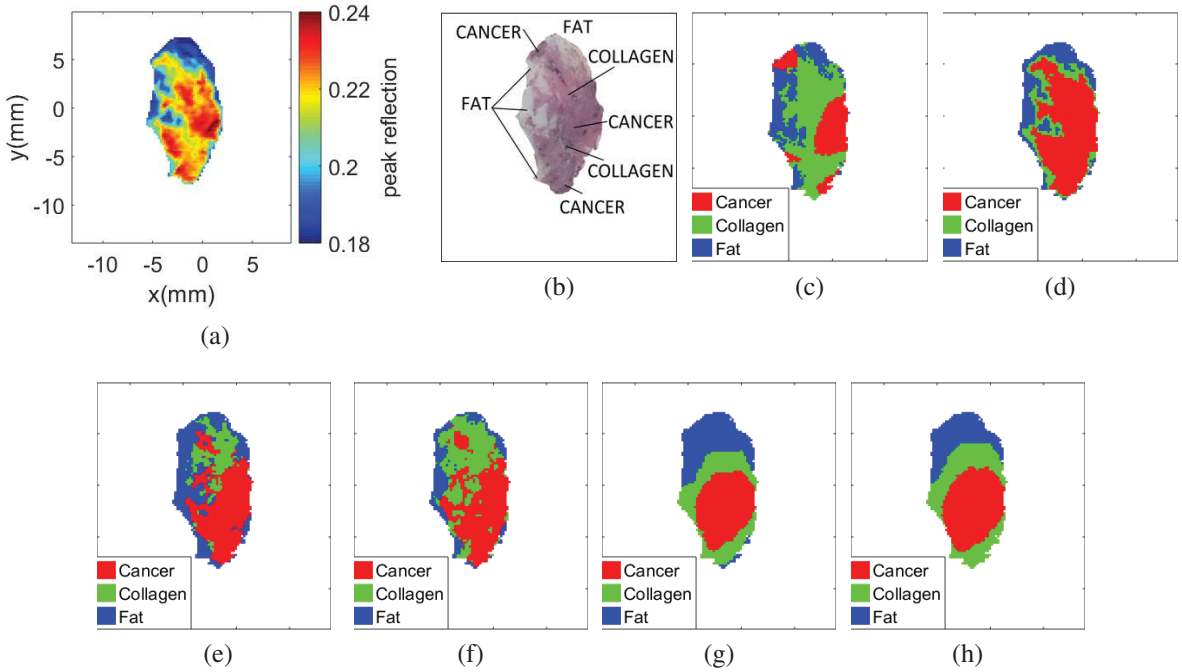


Figure 3.9: Sample ND15588 block. (a) THz image. (b) Pathology image. (c) Morphed Pathology. (d) 1D MCMC model. (e) 6D amplitude MCMC model. (f) 6D amplitude EM model. (g) 2D complex MCMC model. (h) 2D complex EM model.

all the regions.

3.8.2. 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. 3.9a, where each pixel is represented by using the peak reflection of the THz waveform [8]. The dimensions of this block sample are $7.5 \times 14.9mm$ and its thickness is $\sim$3-4mm. 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 is not utilized for this type of samples in the 1D 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. 3.7a, this image was collected from scanning the paraffin block

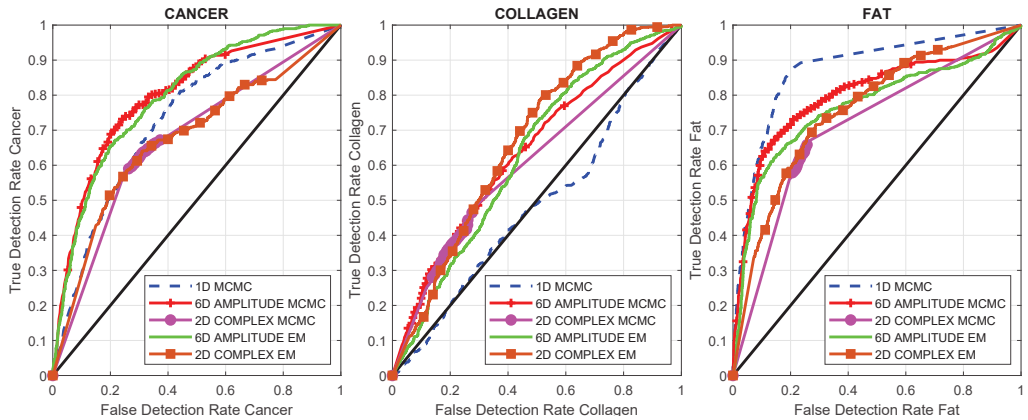


Figure 3.10: ROC curves for sample ND15588 block.

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. 3.7a. 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 Figs. 3.9b and 3.9c, respectively. The classification results obtained by using the 1D MCMC approach[5, 8], 6D MCMC with amplitude spectrum, 6D EM with amplitude spectrum, 2D MCMC with complex spectrum, and 2D EM with complex spectrum are presented in Figs. 3.9d-3.9h, respectively. The corresponding ROC curves are given in Fig. 3.10. The areas underneath the ROC curves are listed in Table 4.1.

Visually the results obtained from the 6D amplitude MCMC and 6D 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 6D 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 1D 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. 3.9d. In terms of the collagen ROC area, the 6D amplitude MCMC and 2D complex EM models achieve the best

Table 3.1: 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

performance among all cases, with that of 2D complex EM being better. However, the 2D complex EM model has a large misclassification of cancer. For the fat region, the 1D MCMC and 6D amplitude MCMC models have the best performance, followed by the 6D amplitude EM model. The 6D 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 6D amplitude EM model. The ROC areas of the 6D 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 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

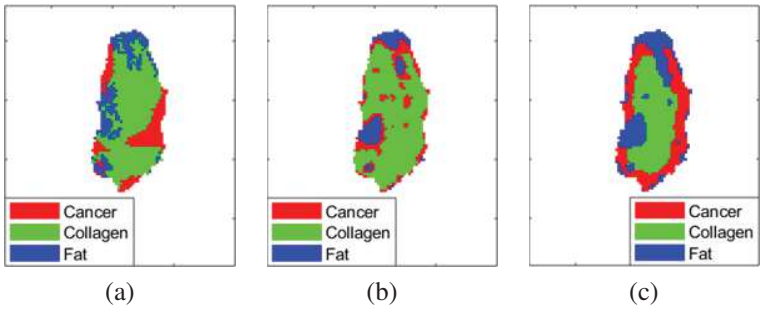


Figure 3.11: Fresh sample ND15588. (a) Morphed pathology. (b) Segmentation results from 1D MCMC. (c) Segmentation results from 2D EM with 8-nearest neighbors.

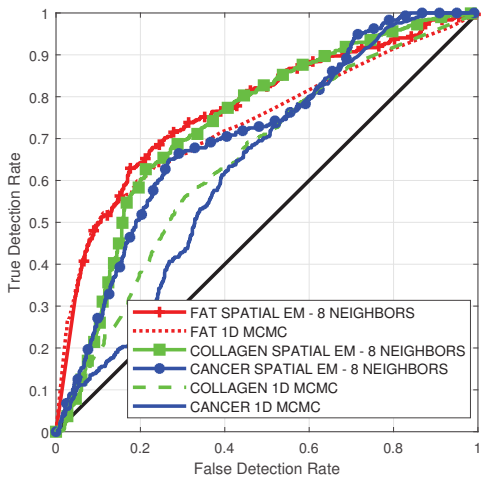


Figure 3.12: ROC curves for fresh sample ND15588.

pathology because both images were taken from the same surface. On the other hand, the THz imaging of fresh samples were taken from different surfaces. Furthermore, the contrast between cancer and healthy non-fatty 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.

Table 3.2: Areas under the ROC curves for sample ND15588 fresh.

Method	Cancer	Collagen	Fat
1D MCMC	0.6338	0.6521	0.7372
Spatial EM - 4 neighbors	0.7092	0.7400	0.7721
Spatial EM - 8 neighbors	0.7099	0.7401	0.7726
Spatial EM - 24 neighbors	0.7126	0.7408	0.7750

Table 3.3: Areas under the ROC curves for sample ND15588 fresh: LOOP vs. 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

3.8.3. Results with Spatial Prior

The results in this section are obtained by applying the newly proposed spatial image segmentation algorithm to sample ND15588. Fig. 3.11a shows the morphed pathology obtained through mesh morphing [5], which represents our ground truth. Fig. 3.11b presents the classification results obtained through a 1D GMM with MCMC as described in [8]. Fig. 3.11c shows the classification results using the 2D spatial EM approach proposed in this chapter. In these figures, we can observe that the spatial model presents better region correlation with the morphed pathology than the 1D MCMC model.

Fig. 3.12 shows the ROC curves of the classification models, which provide the quantitative evaluation of these models in the form of true vs. false detection rate for each region within the tissue. As shown in Fig. 3.12, the spatial model performs better than the 1D MCMC approach for all regions. This can be further confirmed in Table 3.2, which presents the areas under the ROC curves. As shown in Table 3.2, we can observe that the area under these curves slightly increases as the number of neighbors increases.

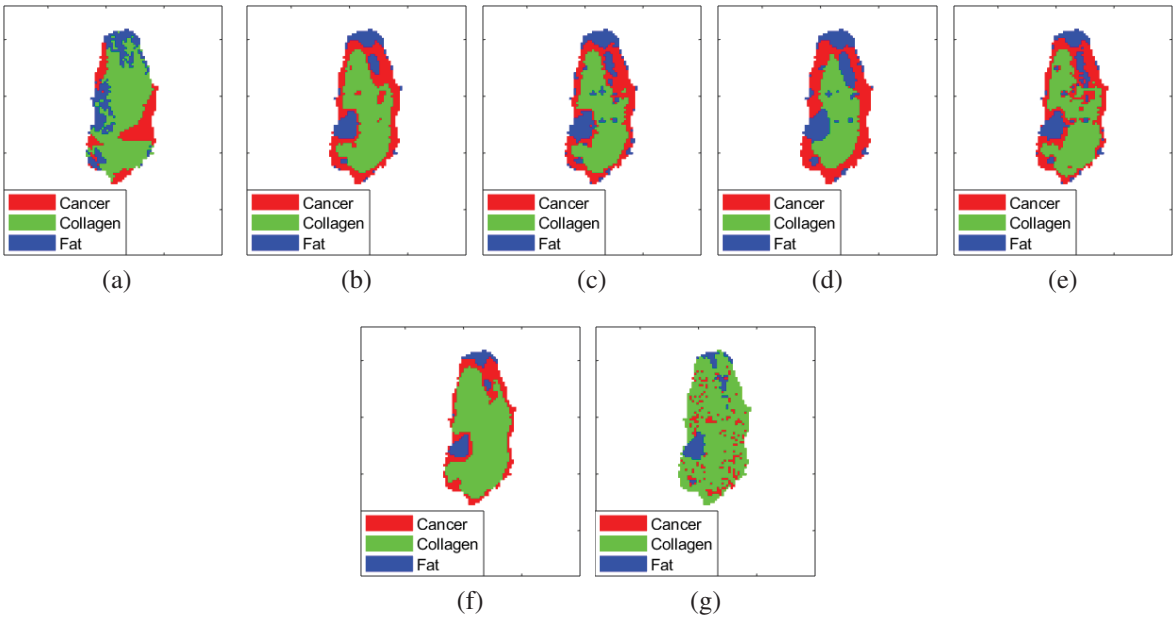


Figure 3.13: Sample ND15588 fresh with different classification methods. (a) Morphed histopathology mask.(b) 2D amplitude MCMC with LOOP. (c) 2D amplitude MCMC with PCA. (d) 3D complex MCMC with LOOP. (e) 3D complex MCMC with PCA. (f) K-means clustering with full spectrum. (g) SVM clustering with full spectrum.

Table 3.4: 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

3.8.4. 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[29], K-means[31], and support vector machine (SVM)[32]. 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 by both PCA and LOOP. The low dimension vectors obtained from PCA or LOOP are further processed by using amplitude or complex MCMC.

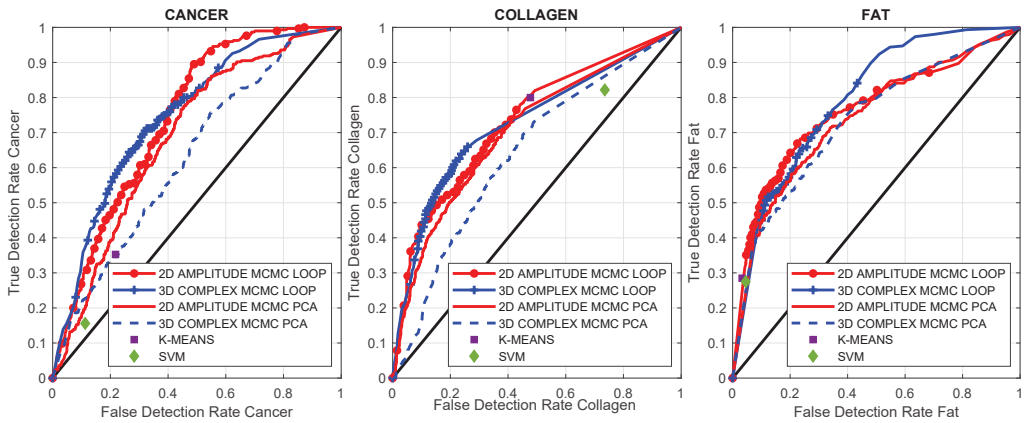


Figure 3.14: ROC curves for sample ND15588 fresh with different classification methods.

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. 3.13a. The classification results of 2D amplitude MCMC with LOOP, 2D amplitude MCMC with PCA, 3D complex MCMC with LOOP, 3D complex MCMC with PCA, K-means, and SVM are shown in Figs. 3.13b-3.13g, respectively. The corresponding ROC curves of the cancer, collagen, and fat regions are shown in Fig. 3.14. The areas underneath the ROC curves are listed in Table 3.3. 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 positive probabilities. Consequently, the results from K-means and SVM are represented as single dots on the ROC curves in Fig. 3.14. The true and false detection rates of K-Means and SVM for all the regions within this sample are shown in Table 3.4.

Overall the 2D amplitude MCMC and the 3D complex MCMC models with LOOP achieves the best performance among all the different methods. K-means achieves comparable results with respect to 2D 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.

3.9. 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 dimension feature vector, which is then modeled by using multivariate GMMs. The low dimension 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 1D 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 1D features in the 1D MCMC algorithm with 2D 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 [8] and [5]. The classification of tumor tissues with 3 or more regions still remains as a significant challenge for future works.

Additionally, the addition of a spatial prior moderately improves the performance of breast cancer detection by exploiting the spatial correlation among neighboring pixels in THz images. Experimental results on freshly excised human tissue have demonstrated that the spatial GMM model achieves better detection rates when compared to 1D MCMC.

3.10. Appendix

3.10.1. Copyright Permission

Copyright Clearance Center RightsLink®

Home | Help | Email Support | Sign in | Create Account

IEEE Requesting permission to reuse content from an IEEE publication

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

Author: Tanny Chavez

Publication: IEEE Transactions on Terahertz Science and Technology

Publisher: IEEE

Date: March 2020

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

Spatial Image Segmentation for Breast Cancer Detection in Terahertz Imaging

Conference Proceedings:
2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting

Author: Tanny Chavez

Publisher: IEEE

Date: 5 July 2020

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

References

[1] World Cancer Research Fund, "Breast cancer statistics," https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics, 2018, [Accessed: May 29, 2019].

[2] Breastcancer.org, "What is lumpectomy?" https://www.breastcancer.org/treatment/surgery/lumpectomy/what_is, March 4, 2015, [Accessed: May 29, 2019].

[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," *Annals of Surgical Oncology*, vol. 26, no. 5, pp. 1238–1244, May 2019. doi: 10.1245/s10434-019-07247-5. [Online]. Available: https://doi.org/10.1245/s10434-019-07247-5

[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. doi: 10.1364/BOE.9.001334. [Online]. Available: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-9-3-1334

[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," *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 39, no. 12, pp. 1283–1302, Dec 2018. doi: 10.1007/s10762-018-0529-8. [Online]. Available: https://doi.org/10.1007/s10762-018-0529-8

[6] N. Vohra, T. Bowman, P. M. Diaz, N. Rajaram, K. Bailey, and M. El-Shenawee, "Pulsed terahertz reflection imaging of tumors in a spontaneous model of breast cancer," *Biomedical Physics & Engineering Express*, vol. 4, no. 6, p. 065025, oct 2018. doi: 10.1088/2057-1976/aae699. [Online]. Available: https://doi.org/10.1088%2F2057-1976%2Faae699

[7] T. Bowman, N. Vohra, K. Bailey, and M. O. El-Shenawee, "Terahertz tomographic imaging of freshly excised human breast tissues," *Journal of Medical Imaging*, vol. 6, no. 2, pp. 1 – 13 – 13, 2019. doi: 10.1117/1.JMI.6.2.023501. [Online]. Available: https://doi.org/10.1117/1.JMI.6.2.023501

[8] T. Bowman, T. Chavez, K. Khan, J. Wu, A. Chakraborty, N. Rajaram, K. Bailey, and M. El-Shenawee, "Pulsed terahertz imaging of breast cancer in freshly excised murine tumors," *Journal of Biomedical Optics*, vol. 23, no. 2, p. 026004, 2018. doi: 10.1117/1.JBO.23.2.026004.

[9] J. Shi, Y. Wang, T. Chen, D. Xu, H. Zhao, L. Chen, C. Yan, L. Tang, Y. He, H. Feng, and J. Yao, "Automatic evaluation of traumatic brain injury based on terahertz imaging with machine learning," *Opt. Express*, vol. 26, no. 5, pp. 6371–6381, Mar 2018. doi: 10.1364/OE.26.006371. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-26-5-6371

[10] Y. Cao, J. Chen, P. Huang, W. Ge, D. Hou, and G. Zhang, "Inspecting human colon adenocarcinoma cell lines by using terahertz time-domain reflection spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 211, pp. 356 – 362, 2019. doi: 10.1016/j.saa.2018.12.023. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1386142518310850

[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," *Journal of Biomedical Optics*, vol. 23, no. 4, pp. 1 – 8 – 8, 2018. doi: 10.1117/1.JBO.23.4.045001. [Online]. Available: https://doi.org/10.1117/1.JBO.23.4.045001

[12] H. Liu, Z. Zhang, X. Zhang, Y. Yang, Z. Zhang, X. Liu, F. Wang, Y. Han, and C. Zhang, "Dimensionality reduction for identification of hepatic tumor samples based on terahertz time-domain spectroscopy," *IEEE Transactions on Terahertz Science and Technology*, vol. 8, no. 3, pp. 271–277, May 2018. doi: 10.1109/TTHZ.2018.2813085.

[13] Z. Zhang, H. Liu, and C. Zhang, "Terahertz pulse data dimensional reduction and classification for hepatic tissue samples," in *2018 43rd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)*, Sep. 2018. ISSN 2162-2035 pp. 1–3. doi: 10.1109/IRMMW-THz.2018.8510381.

[14] M. W. Ayech and D. Ziou, "Segmentation of terahertz imaging using k-means clustering based on ranked set sampling," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2959 – 2974, 2015. doi: https://doi.org/10.1016/j.eswa.2014.11.050. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417414007490

[15] X. Yin, W. Mo, Q. Wang, and B. Qin, "A terahertz spectroscopy nondestructive identification method for rubber based on CS-SVM," *Advances in Condensed Matter Physics*, vol. 2018, no. 1618750, pp. 1–8, 2018. doi: 10.1155/2018/1618750. [Online]. Available: https://doi.org/10.1155/2018/1618750

[16] Y. Li, X. A. Shen, R. L. Ewing, and J. Li, "Terahertz spectroscopic material identification using approximate entropy and deep neural network," in *2017 IEEE National Aerospace and Electronics Conference (NAECON)*, June 2017. ISSN 2379-2027 pp. 52–56. doi: 10.1109/NAECON.2017.8268744.

[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 *2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting*, July 2019. ISSN 1522-3965 pp. 525–526. doi: 10.1109/APUSNCURSINRSM.2019.8888653

[18] D. Reynolds, *Gaussian Mixture Models*. Boston, MA: Springer US, 2015, pp. 827–832. ISBN 978-1-4899-7488-4. [Online]. Available: https://doi.org/10.1007/978-1-4899-7488-4_196

[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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 3, pp. 813–819, March 2014. doi: 10.1109/JSTARS.2014.2305981.

[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 *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2013. ISSN 1094-687X pp. 4326–4329. doi: 10.1109/EMBC.2013.6610503.

[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 Transactions on Geoscience and Remote Sensing*, vol. 32, no. 5, pp. 1087–1095, Sep. 1994. doi:

10.1109/36.312897.

[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, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 236–243.

[23] D. Maiwald and D. Kraus, "Calculation of moments of complex wishart and complex inverse wishart distributed matrices," *IEE Proceedings - Radar, Sonar and Navigation*, vol. 147, no. 4, pp. 162–168, Aug 2000. doi: 10.1049/ip-rsn:20000493.

[24] I. Yildirim, "Bayesian inference: Gibbs sampling," *Technical Note, University of Rochester*, 2012.

[25] I. Alvarez, J. Niemi, and M. Simpson, "Bayesian inference for a covariance matrix," in *26th Annual Conference on Applied Statistics in Agriculture*, April 27-29, 2014. doi: 10.4148/2475-7772.1004.

[26] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[27] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 3, pp. 809–823, March 2012. doi: 10.1109/TGRS.2011.2162649

[28] 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

[29] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 8 2010. doi: 10.1002/wics.101. [Online]. Available: https://doi.org/10.1002/wics.101

[30] M. El-Shenawee, N. Vohra, T. Bowman, and K. Bailey, "Cancer detection in excised breast tumors using terahertz imaging and spectroscopy," *Biomedical Spectroscopy and Imaging*,

vol. 8, no. 1-2, pp. 1–9, July 2019. doi: 10.3233/BSI-190187.

[31] S. Lloyd, “Least squares quantization in PCM,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, March 1982. doi: 10.1109/TIT.1982.1056489.

[32] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep 1995. doi: 10.1007/BF00994018. [Online]. Available: https://doi.org/10.1007/BF00994018

Chapter 4

Supervised Bayesian Learning for Breast Cancer Detection in Terahertz Imaging

Tanny Chavez, Nagma Vohra, Keith Bailey, Magda El-Shenawee, and Jingxian Wu

T. Chavez, N. Vohra, K. Bailey, M. El-Shenawee, and J. Wu, "Supervised Bayesian Learning for Breast Cancer Detection in Terahertz Imaging", submitted for publication to *Biomedical Signal Processing and Control* and in review.

4.1. Abstract

This paper proposes a supervised multinomial Bayesian learning algorithm for breast cancer detection by using terahertz (THz) images 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 require much less training data when compared to 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 will negatively impact 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 within THz images of freshly excised breast cancer samples.

4.2. Introduction

Breast cancer is one of the most common forms of cancer in women across the U.S., with approximately 1 in 8 women estimated to be diagnosed with breast cancer during their lifetime [1]. Among feasible treatment options for early detected breast cancer, mastectomies and breast conserving surgery (BCS) correspond to the customary care approaches. For instance, BCS removes the cancerous tumor surrounded by a small margin of healthy breast tissue. 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 oscillates between 20-30% [2]. Therefore, it is necessary to develop an imaging benchmark for the detection of breast cancer within freshly excised BCS samples such that the surgeon can evaluate the margins of the tissue in the operating room and reduce its overall re-excision rates.

Terahertz (THz) imaging has shown great potentials for material characterization in a vast variety of applications, such as integrated circuit inspection [3], security screening [4], food inspection [5], and biomedical imaging [6–11]. The common objective across these studies is the classification of the reflected THz pulse into a fixed number of categories, but with different segmentation technique based on unsupervised or supervised learning methods. In general, unsupervised learning algorithms, such as mixture models [6, 12], and Fuzzy C-means [11], make inferences on patterns among the input observations without utilizing a training stage. These techniques are useful for initial data 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 image segmentation include support vector machine (SVM) [8, 9, 13], partial least squares-discriminant analysis (PLS-DA) [9], K-nearest neighbors [8, 13], random forest [8, 14], and convolutional neural networks (CNN) [3, 4, 15]. Although supervised learning al-

gorithms 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 [10] 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) [9, 16], and the low-dimensional ordered orthogonal projection (LOOP) [6, 7] algorithm. Once the most relevant features have been identified, the segmentation algorithm utilizes these attributes to perform inferences on the parameters of their discriminating models.

This chapter introduces a novel supervised image segmentation algorithm for the detection of breast cancer in THz images 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 [7, 17] or binary classifications [18]. Two non-linear regression models, polynomial regression and kernel regression with random Fourier features (RFF) [19], 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, 20], this chapter 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 FFPE process of the tissue. Due to dehydration during the FFPE 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 [21]. 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 chapter is organized as follows. Section 4.3 introduces the THz system and the procedure to collect the images. Section 4.4 presents the proposed regression model, and its training and testing procedures. Section 4.5 shows the experimental results. Section 4.6 concludes this study.

4.3. Materials and Methods

The experimental set-up was established in Dr. El-Shenawee's Terahertz Imaging and Spectroscopy Lab at the University of Arkansas. The raw experimental data was provided by Ms.

Nagma Vohra, PhD candidate in Dr. El-Shenawee's group.

4.4. Theory and Algorithm

4.4.1. 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 $\boldsymbol{\mathcal{V}} \in \mathcal{R}^{N_1 \times N_2 \times F}$, 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 $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_{N_s}]$, where $\mathbf{v}_n \in \mathcal{R}^F$ represents the amplitude of the frequency domain spectrum of the reflected waveform in the n-th pixel, and $n = \{1, \ldots, N_s\}$ 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 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 are 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 $\mathbf{x}_n \in \mathcal{R}^{1 \times L}$, where $n = \{1, \ldots, N\}$ 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 4.4.3.

4.4.2. 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, $\mathbf{z} \in \mathcal{R}^N$, for non-binary partitions of the dataset [17].

Given the estimated value of the latent variable, and a set of estimated thresholds, $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \ldots, \alpha_K\}$, 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.

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

$$z_n \overset{\text{ind}}{\sim} \mathcal{N}\left(\mathbf{w}_n \boldsymbol{\beta}, \sigma^2\right), \tag{4.1}$$

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

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 $h : \mathbf{x}_n \in \mathcal{R}^L \rightarrow h(\mathbf{x}_n) \in \mathcal{R}^U$ represents the feature mapping, and $U > L$. 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

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

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

$$\mathcal{K}(\mathbf{x}_m, \mathbf{x}_n) = e^{-\nu||\mathbf{x}_m - \mathbf{x}_n||^2} \tag{4.2}$$

where ν is the kernel parameter.

The complexity of the kernel regression model increases with the size of the training dataset. The number of training samples used in this study is in general much smaller compared to 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 [19], 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 $h_{\text{RFF}} : \mathbf{x}_n \in \mathcal{R}^L \rightarrow h(\mathbf{x}_n) \in \mathcal{R}^V$ with $V < U$ and

$$\mathcal{K}(\mathbf{x}_m, \mathbf{x}_n) \approx h_{\text{RFF}}(\mathbf{x}_m)h_{\text{RFF}}(\mathbf{x}_n)^T. \tag{4.3}$$

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

theorem as

$$\mathcal{K}(\mathbf{x}_m - \mathbf{x}_n) = \int_{\mathcal{R}^L} e^{i\boldsymbol{\omega}^T(\mathbf{x}_m - \mathbf{x}_n)} P(\boldsymbol{\omega}) d\boldsymbol{\omega} \tag{4.4}$$

where $P(\boldsymbol{\omega})$ corresponds to the Fourier transform of the kernel, and $\boldsymbol{\omega} \in \mathcal{R}^{L\times 1}$ is the vector corresponding to the frequency domain variable. Since it is not possible to directly compute (4.4), 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 $P(\boldsymbol{\omega}) = \mathcal{N}(\mathbf{0}_L, 2\nu\mathbf{I}_L)$. By following the Monte Carlo approach, the kernel function in (4.4) can be approximated by

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

where $\boldsymbol{\omega}_q \overset{\text{iid}}{\sim} P(\boldsymbol{\omega})$, and Q is the total number of Monte Carlo iterations [19]. Through this expression, the feature space defined by RFF can then be expressed as

$$h_{\text{RFF}}(\mathbf{x}) = \frac{1}{\sqrt{Q}} \begin{bmatrix} \cos(\boldsymbol{\Omega}^T \mathbf{x}) \\ \sin(\boldsymbol{\Omega}^T \mathbf{x}) \end{bmatrix} \in \mathcal{R}^{2Q\times 1} \tag{4.5}$$

where $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_Q] \in \mathcal{R}^{L\times Q}$. In (4.5), 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, $Q = \sqrt{N}\log(N)$ [19].

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

$$z_n \overset{\text{ind}}{\sim} \mathcal{N}\left(\mathbf{w}_n \boldsymbol{\beta}, \sigma^2\right), \tag{4.6}$$

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

4.4.3. 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 an Markov chain Monte Carlo (MCMC) method.

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 pathologist's 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 [21]. 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 are 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 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 dataset only if the probability exceeds a certain threshold, and the corresponding region matches the pathology results. In this chapter, 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.

Parameter initialization

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

To ensure that the $\boldsymbol{\alpha}$ parameter covers the entire latent variable domain, $\mathcal{R}$, certain elements within this parameter are manually fixed as $\alpha_0 = -\infty$, $\alpha_1 = 0$, and, $\alpha_K = \infty$ [17]. 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,

$$\mathbf{q} = -\mathbf{W}\boldsymbol{\beta}, \tag{4.7}$$

where $\mathbf{W} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \cdots, \mathbf{w}_N^T]^T$, $\mathbf{q} = [q_1, \cdots, q_N]^T \in \mathcal{R}^{N\times 1}$ 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

$$\boldsymbol{\beta} = -(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{q}. \tag{4.8}$$

In the ground truth data 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 the exact results in (4.8) 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 chapter we choose $\epsilon = 0.0013$.

Similarly 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, $\{\alpha_2, \ldots, \alpha_{K-1}\}$. 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,

$$\alpha_{K-1} = \frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{w}_n\boldsymbol{\beta} + \Phi^{-1}[1 - \Pr(y_n = K)]\right\}. \tag{4.9}$$

Since this chapter 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}.

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 a MCMC process. The prior distributions of the model parameters $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ are defined as:

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

with $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$ representing the hyper-parameters of this approach. In this chapter, 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 [17]:

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

$$z_n|\boldsymbol{\beta}, \boldsymbol{\alpha}, y_n = k \sim \begin{cases} 0 & ; z_n \leq \alpha_{k-1} \\ \dfrac{\phi\left(\mathbf{w}_n\boldsymbol{\beta}, \sigma^2; z_n\right)}{\Phi\left(\frac{\alpha_k - \mathbf{w}_n\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{\alpha_{k-1} - \mathbf{w}_n\boldsymbol{\beta}}{\sigma}\right)} & ; \alpha_{k-1} < z_n < \alpha_k \\ 0 & ; z_n \geq \alpha_k \end{cases} \tag{4.10}$$

  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 $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}|\mathbf{z} \sim \mathcal{N}(\mu_\beta, \boldsymbol{\Sigma}_\beta), \tag{4.11}$$

  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 $\boldsymbol{\alpha}$,

$$\alpha_k|\mathbf{z}, \mathbf{y}, \alpha_{j\neq k} \sim \mathcal{U}(a, b), \tag{4.12}$$

  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 4, 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.

Algorithm 4 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 (4.8) and (4.9), respectively
for $j = 1, \ldots, 10M$ **do**
  Draw $\mathbf{Z}^{(j)}$ from (4.10) using $\boldsymbol{\beta}^{(j-1)}$, $\boldsymbol{\alpha}^{(j-1)}$, and $\mathbf{y}$.
  Draw $\boldsymbol{\beta}^{(j)}$ from (4.11) using $\mathbf{Z}^{(j)}$.
  Draw the unknown elements within $\boldsymbol{\alpha}^{(j)}$ from (4.12) 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 $[\boldsymbol{\beta}^{(i)}, \boldsymbol{\alpha}^{(i)}]_{i=1}^{M}$.

4.4.4. 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 go under the same pre-processing procedures, which include obtaining the frequency response of the pulse and dimension reduction.

Once the corresponding model parameters are obtained during the training phase, as described in section 4.4.3, the region assignment is performed by using a 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

$$z_n^{(i)} \sim \mathcal{N}(\mathbf{w}_n \boldsymbol{\beta}^{(i)}, \sigma^2), \quad \text{for } i = 1, \cdots, M \tag{4.13}$$

Thus

$$\Pr(\alpha_{k-1}^{(i)} < z_n^{(i)} \leq \alpha_{k-1}^{(i)}) = \Phi\left(\frac{\alpha_k^{(i)} - \mathbf{w}_n \boldsymbol{\beta}^{(i)}}{\sigma}\right) - \Phi\left(\frac{\alpha_{k-1}^{(i)} - \mathbf{w}_n \boldsymbol{\beta}^{(i)}}{\sigma}\right) \tag{4.14}$$

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

Table 4.1: 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

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

$$\Pr(y_n = k) = \frac{1}{M} \sum_{i=1}^{M} \left[\Pr(\alpha_{k-1}^{(i)} < z_n^{(i)} \leq \alpha_{k-1}^{(i)}) \right] \tag{4.15}$$

where M is the total number of stored MCMC iterations. With (4.15), we evaluate the likelihood of each pixel from the testing data with respect to every region in the tissue.

4.5. Experimental results

The experiment results are obtained by applying the proposed multinomial probit regression algorithm on the testing data. The training and testing data are from freshly excised xenograft murine samples with 3 regions each, such as cancer, muscle or fibro, and fat. These 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. The results from the proposed algorithms are compared to two previously published unsupervised learning approaches based on

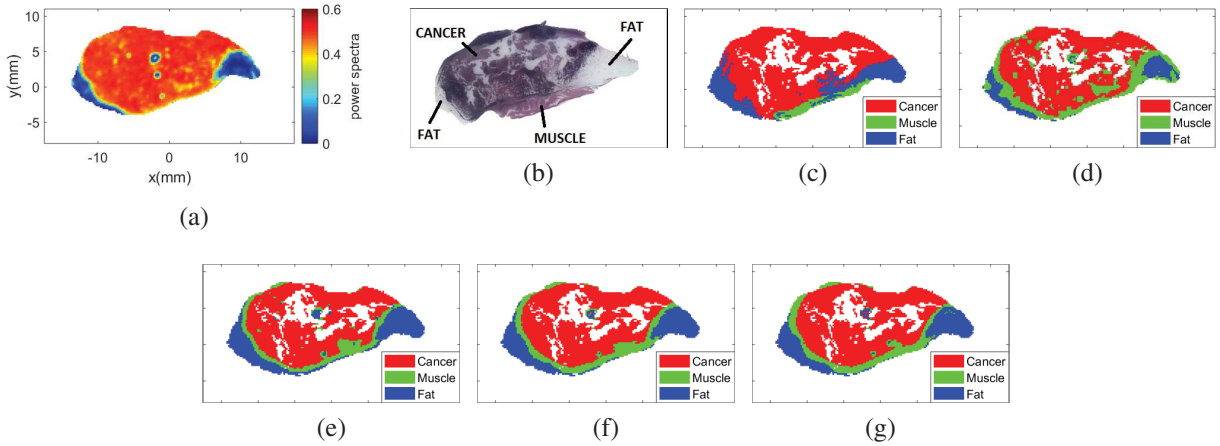


Figure 4.1: Sample Mouse 9B Fresh. (a) THz image [24]. (b) Pathology image [24]. (c) Morphed Pathology [24]. (d) 1D MCMC model [24]. (e) 2D unsupervised EM model. (f) 3D supervised polynomial regression model (this work). (g) 3D supervised RFF kernel model (this work).

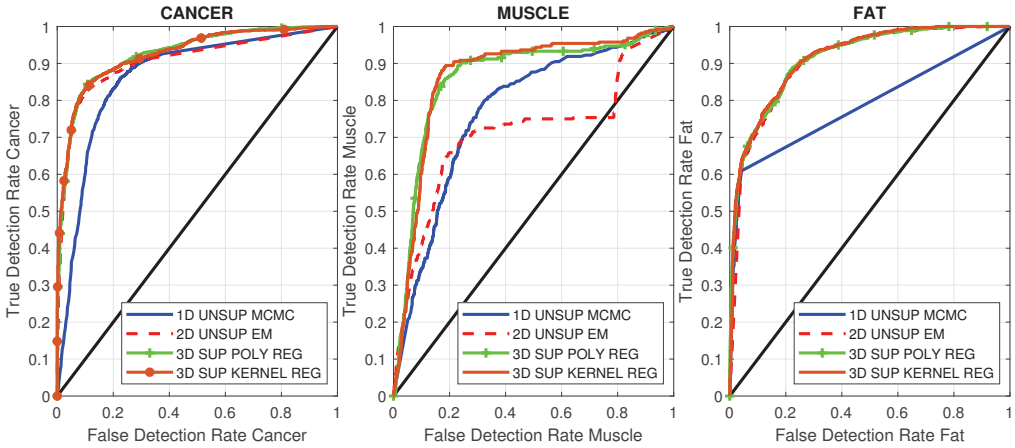


Figure 4.2: ROC curves for sample Mouse 9B Fresh.

GMM, which are 1-dimensional (1D) MCMC [22] and 2-dimensional (2D) EM [6]. Source codes for the multinomial probit regression algorithm can be found in [23]. 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 4.7.1.

4.5.1. 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. 4.1a, 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 region is not so obvious. This could be because the electrical properties of muscle and cancer are identical in the THz range [22]. Fig. 4.1b represents the pathology analysis of this sample, which clearly indicates the location and the extent of the regions within the tissue. Fig. 4.1c shows the morphed pathology results obtained from the mesh morphing algorithm [21]. Figs. 4.1d and 4.1e correspond to the 1D MCMC [22] and 2D EM [6] segmentation results, respectively. Finally, Figs. 4.1f and 4.1g represent the multinomial probit segmentation results obtained by using the 3D polynomial and kernel regression models introduced in this chapter, 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 $Q = 20$.

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. 4.2. 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 gains over the two unsupervised algorithms.

This performance gain can be quantified by analyzing the areas under the ROC curves, and the results are in Table 4.1. 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 chapter obtain the largest areas under the ROC curves for all regions, with muscle representing the highest performance gain from 71.35% to 86.80%.

4.5.2. 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. 4.3a 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. 4.3b and 4.3c correspond to the histopathology analysis of the tissue and its corresponding morphed mask. Figs. 4.3d and 4.3e represent the results obtained through the unsupervised Gaussian mixture models. The linear and kernel regression models are represented in Figs. 4.3f and 4.3g, 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 $Q = 20$.

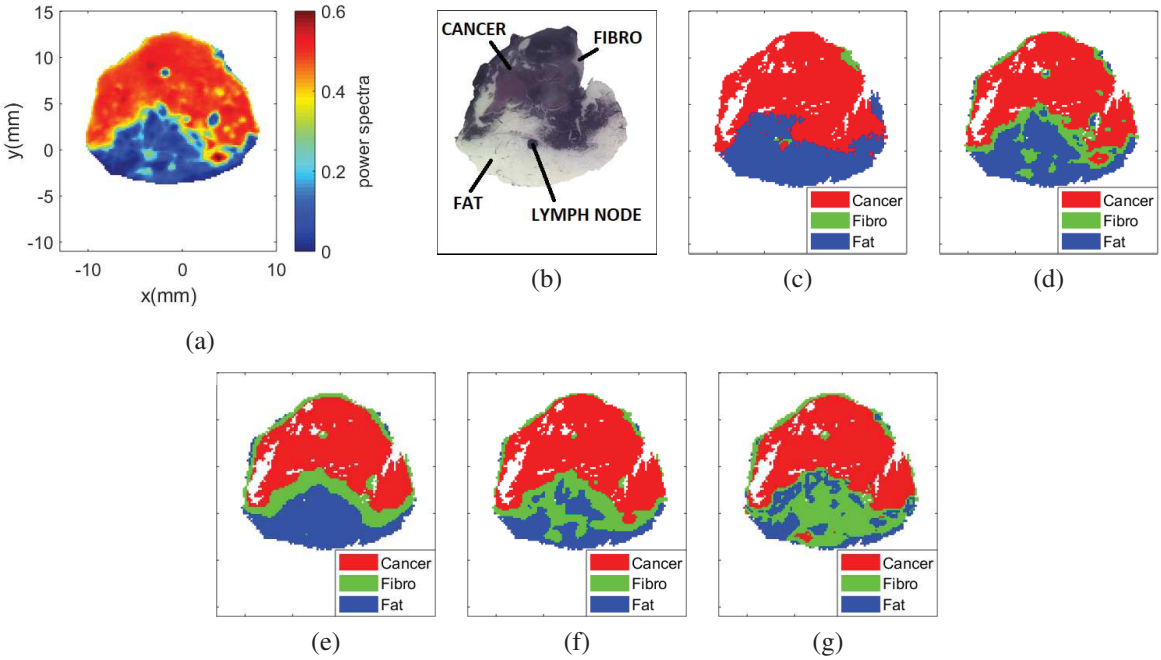


Figure 4.3: Sample Mouse 13A Fresh. (a) THz image [21]. (b) Pathology image [21]. (c) Morphed Pathology [21]. (d) 1D MCMC model [21]. (e) 2D unsupervised EM model. (f) 2D supervised linear regression model (this work). (g) 2D supervised RFF kernel model (this work).

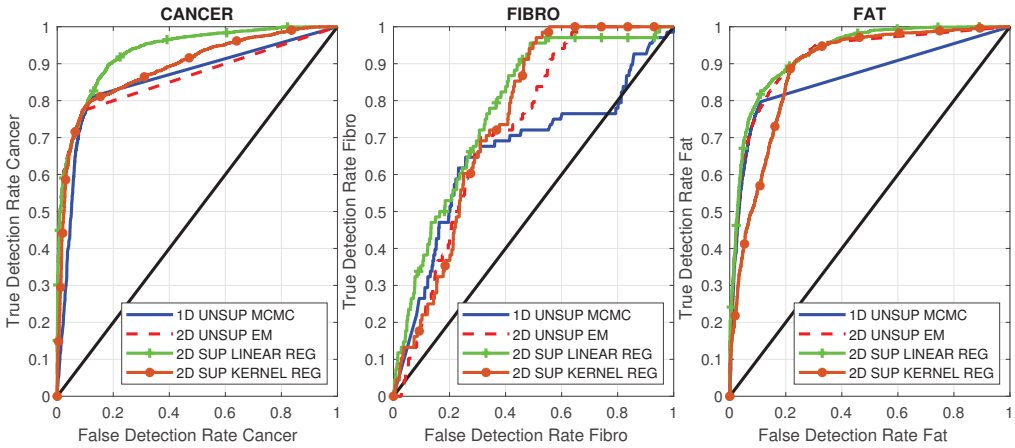


Figure 4.4: ROC curves for sample Mouse 13A Fresh.

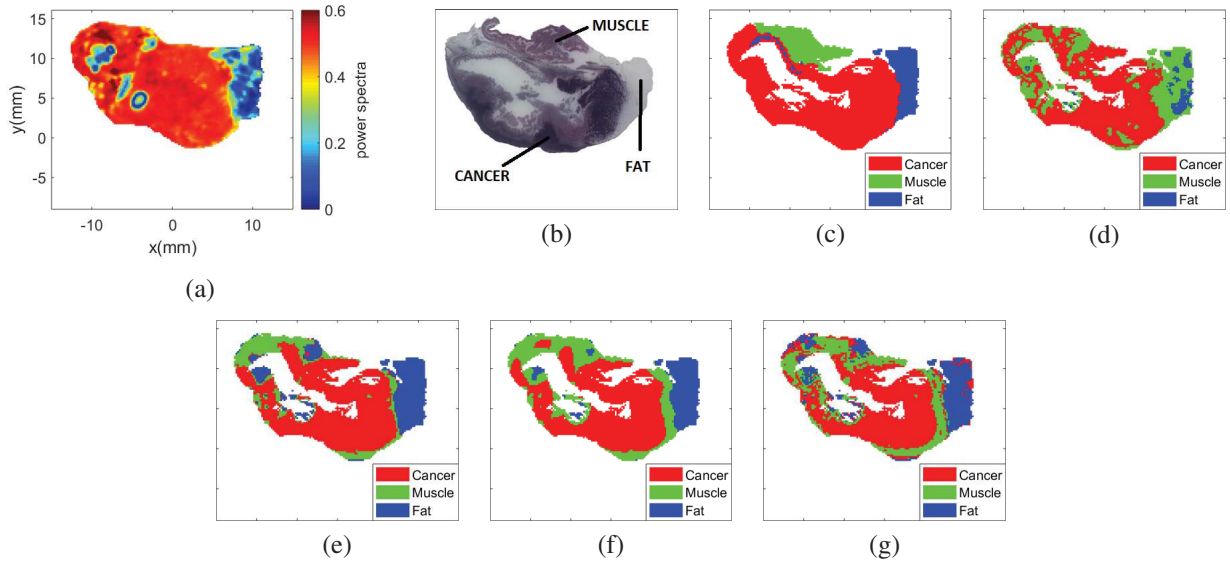


Figure 4.5: Sample Mouse 10B Fresh. (a) THz image. (b) Pathology image. (c) Morphed Pathology. (d) 1D MCMC model. (e) 2D unsupervised EM model. (f) 2D supervised linear regression model (this work). (g) 3D supervised RFF kernel model (this work).

The ROC curves of the classifiers are shown in Fig. 4.4, 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 4.1, 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%.

4.5.3. Mouse 10B Fresh

Finally, the third sample is mouse 10B fresh, which contains 3 regions: cancer, muscle, and fat. Fig. 4.5a represents the THz image of this sample. Figs. 4.5b and 4.5c 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. 4.5a. Figs. 4.5d and 4.5e represent the unsupervised classification results obtained through the 1D MCMC and 2D EM approaches, respectively. Figs. 4.5f and 4.5g illustrate the segmentation results obtained through

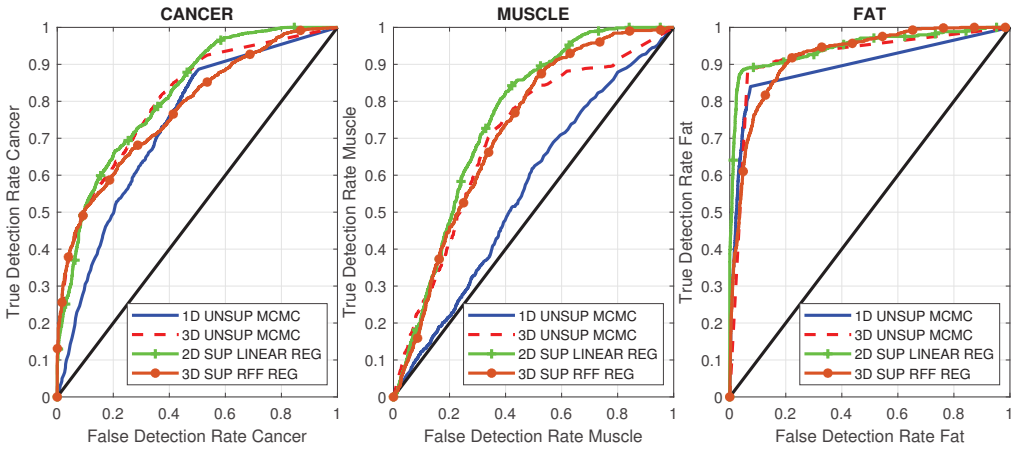


Figure 4.6: ROC curves for sample Mouse 10B Fresh.

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 $Q = N\log(N) = 442$.

The quantitative evaluation of the results are shown in Fig. 4.6 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 4.1, 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.

4.6. Conclusions

We have proposed a supervised multinomial Bayesian learning method for cancer detection by using THz images of freshly excised BCS 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, have been 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 analysis, 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 to 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 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.

4.7. Appendix

4.7.1. 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 $P(y_n = k) \geq \delta$. 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] American Cancer Society, *Breast Cancer Facts & Figures 2019-2020*. Atlanta: American Cancer Society, Inc., 2019. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/

[2] L. C. Elmore and J. A. Margenthaler, "A tale of two operations: re-excision as a quality measure," *Gland Surgery*, vol. 8, no. 6, 2019.

[3] Q. Mao, Y. Zhu, C. Lv, Y. Lu, X. Yan, S. Yan, and J. Liu, "Convolutional neural network model based on terahertz imaging for integrated circuit defect detections," *Optics Express*, vol. 28, no. 4, pp. 5000–5012, 2020.

[4] X. Yang, T. Wu, L. Zhang, D. Yang, N. Wang, B. Song, and X. Gao, "CNN with spatio-temporal information for fast suspicious object detection and recognition in thz security images," *Signal Processing*, vol. 160, pp. 202 – 214, 2019. doi: https://doi.org/10.1016/j.sigpro.2019.02.029. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168419300866

[5] W. Liu, C. Liu, X. Hu, J. Yang, and L. Zheng, "Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics," *Food Chemistry*, vol. 210, pp. 415 – 421, 2016. doi: https://doi.org/10.1016/j.foodchem.2016.04.117. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0308814616306458

[6] T. Chavez, N. Vohra, J. Wu, K. Bailey, and M. El-Shenawee, "Breast cancer detection with low-dimension ordered orthogonal projection in terahertz imaging," *IEEE Transactions on Terahertz Science and Technology*, pp. 1–1, 2019. doi: 10.1109/TTHZ.2019.2962116

[7] T. Chavez, N. Vohra, J. Wu, N. Rajaram, K. Bailey, and 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*, 2020.

[8] W. Liu, R. Zhang, Y. Lu, R. She, K. Zhou, B. Fang, G. Wei, and G. Li, "Classification of terahertz pulsed signals from breast tissues using wavelet packet energy feature exaction and machine learning classifiers," in *Infrared, Millimeter-Wave, and Terahertz Technologies VI*, C. Zhang, X.-C. Zhang, and M. Tani, Eds., vol. 11196, International Society for Optics and Photonics. SPIE, 2019, pp. 15 – 22. doi: 10.1117/12.2537277. [Online]. Available: https://doi.org/10.1117/12.2537277

[9] N. Qi, Z. Zhang, Y. Xiang, Y. Yang, X. Liang, and 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*, vol. 7, pp. 2333–2338, 2015. doi: 10.1039/C4AY02665A. [Online]. Available: http://dx.doi.org/10.1039/C4AY02665A

[10] F. Wahaia, I. Kašalynas, L. Minkevičius, C. C. Silva, A. Urbanowicz, and G. Valušis, "Terahertz spectroscopy and imaging for gastric cancer diagnosis," *Journal of Spectral Imaging*, vol. 9, no. 1, p. a2, 2020. doi: 10.1255/jsi.2020.a2. [Online]. Available: https://doi.org/10.1255/jsi.2020.a2

[11] Y. Wang, Z. Sun, D. Xu, L. Wu, J. Chang, L. Tang, Z. Jiang, B. Jiang, G. Wang, T. Chen, H. Feng, and J. Yao, "A hybrid method based region of interest segmentation for continuous wave terahertz imaging," *Journal of Physics D: Applied Physics*, vol. 53, no. 9, p. 095403, dec 2019. doi: 10.1088/1361-6463/ab58b6. [Online]. Available: https://doi.org/10.1088%2F1361-6463%2Fab58b6

[12] U. R. Acharya, F. Molinari, S. V. Sree, S. Chattopadhyay, K.-H. Ng, and J. S. Suri, "Automated diagnosis of epileptic EEG using entropies," *Biomedical Signal Processing and Control*, vol. 7, no. 4, pp. 401 – 408, 2012. doi: https://doi.org/10.1016/j.bspc.2011.07.007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809411000838

[13] H. Kalbkhani, M. G. Shayesteh, and B. Zali-Vargahan, "Robust algorithm for brain magnetic resonance image (MRI) classification based on garch variances series," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 909 – 919, 2013. doi: https://doi.org/10.1016/j.bspc.2013.09.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809413001262

[14] S. Mishra, B. Majhi, P. K. Sa, and L. Sharma, "Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection," *Biomedical Signal Processing and Control*, vol. 33, pp. 272 – 280, 2017. doi: https://doi.org/10.1016/j.bspc.2016.11.021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809416302117

[15] A. Feng-Ping and L. Zhi-Wen, "Medical image segmentation algorithm based on feedback mechanism convolutional neural network," *Biomedical Signal Processing and Control*, vol. 53, p. 101589, 2019. doi: https://doi.org/10.1016/j.bspc.2019.101589. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809419301697

[16] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and discrete wavelet transform," *Biomedical Signal Processing and Control*, vol. 8, no. 5, pp. 437 – 448, 2013. doi: https://doi.org/10.1016/j.bspc.2013.01.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809413000062

[17] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993. [Online]. Available: http://www.jstor.org/stable/2290350

[18] S. S. Chand and K. J. E. Walsh, "Modeling Seasonal Tropical Cyclone Activity in the Fiji Region as a Binary Classification Problem," *Journal of Climate*, vol. 25, no. 14, pp. 5057–5071, 07 2012. doi: 10.1175/JCLI-D-11-00507.1. [Online]. Available: https://doi.org/10.1175/JCLI-D-11-00507.1

[19] P. Milton, H. Coupland, E. Giorgi, and S. Bhatt, "Spatial analysis made easy with linear regression and kernels," *Epidemics*, vol. 29, p. 100362, 2019. doi: https://doi.org/10.1016/j.epidem.2019.100362. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1755436519300337

[20] D. Biswas, A. Gorey, G. C. Chen, S. Vasudevan, N. Sharma, P. Bhagat, and S. Phatak, "Empirical wavelet transform based photoacoustic spectral response technique for assessment of ex-vivo breast biopsy tissues," *Biomedical Signal Processing and Control*, vol. 51, pp. 355 – 363, 2019. doi: https://doi.org/10.1016/j.bspc.2019.02.019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809419300631

[21] T. Chavez, T. Bowman, J. Wu, K. Bailey, and M. El-Shenawee, "Assessment of terahertz imaging for excised breast cancer tumors with image morphing," *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 39, no. 12, pp. 1283–1302, Dec 2018. doi:

10.1007/s10762-018-0529-8

[22] T. Bowman, T. Chavez, K. Khan, J. Wu, A. Chakraborty, N. Rajaram, K. Bailey, and M. El-Shenawee, "Pulsed terahertz imaging of breast cancer in freshly excised murine tumors," *Journal of Biomedical Optics*, vol. 23, no. 2, p. 026004, 2018. doi: 10.1117/1.JBO.23.2.026004.

[23] T. Chavez, N. Vohra, M. El-Shenawee, K. Bailey, and J. Wu, "Source code for "multinomial probit regression for breast cancer detection in terahertz imaging"," https://github.com/taxe10/Multinomial-Probit-Regression, 2020.

[24] T. Chavez, T. Bowman, J. Wu, M. El-Shenawee, and 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*, July 2019. ISSN 1522-3965 pp. 525–526. doi: 10.1109/APUSNCURSINRSM.2019.8888653

Chapter 5

Conclusions

This chapter outlines the contributions of this study and presents a list of tentative guidelines for future directions of this research topic.

5.1. Contributions

This dissertation focuses on the design and implementation of image segmentation approaches for the detection of breast cancer in THz imaging. These algorithms employ a wide array of statistical and machine learning techniques that address the challenges in THz signals analysis, such as the curse of dimensionality, ground truth reliability, and the presence of mixed pixels. The contributions of this dissertation are listed as follows.

First, this study analyzed the importance of the feature selection processes to lessen the impact of the curse of dimensionality. In chapter 2, we introduced the use of 2 physical characteristics, the time-domain peak and the power spectra per pixel, as discriminating features in the classification process [1]. Alternatively, chapter 3 studied the utilization of dimension reduction techniques to systematically identify the most relevant features in the THz waveforms while minimizing the loss of information [2]. Most importantly, this chapter introduced the implementation of a novel dimension reduction technique, LOOP, which employs a modified Gram-Schmidt process to estimate the ordered orthonormal basis of the fundamental features. Experimental results demonstrated that LOOP achieved the best region detection performance when compared to PCA and physical characteristics in human fresh cancer samples [2].

Second, this dissertation evaluated the performance of 3 data representation models: mixture model, MRF, and probit regression. Among the first, chapter 2 studied the implementation of Gaussian and t-distribution mixture models under an unsupervised learning perspective [1]. Furthermore, we explored an alternative definition of the GMM model by incorporating multivariate analysis and spatial probability priors, respectively [2, 3]. In chapter 4, the performance of poly-

nomial and kernel probit regression models was analyzed under a supervised learning regimen [4]. The model parameters across these studies were estimated through unsupervised and supervised MCMC and EM approaches. Experimental results in chapter 4 show that the supervised polynomial regression technique achieved the best region detection performance, followed by unsupervised multivariate GMM with EM within freshly excised xenograft mice tissue.

Third, the proposed approaches employed a soft-clustering technique, which defines a probability driven label assignment procedure. Unlike hard-clustering algorithms, such as K-means and SVM [5, 6], the results obtained through soft-clustering classifiers are not fixed to a single value and can be further tuned to achieve certain detection rates. As shown in chapter 3, GMM obtained higher detection rates when compared to K-means and SVM in THz images of fresh human samples [2].

Fourth, this work introduced a novel reliability based training dataset selection approach to tackle the ground truth reliability challenge in THz imaging [4]. Although the morphing algorithm corrects the tissue deformation discrepancy due to the dehydration of the sample, it is necessary to quantify the accuracy of the ground truth label assignment per pixel before it's usage in the training procedure. Hence, chapter 4 proposes the implementation of an unsupervised EM classifier to measure the ground truth reliability of the training observations prior the parameters estimation procedure. Results on freshly excised murine samples demonstrate that the proposed approach efficiently identifies the most suitable training pixels for the optimal performance of the segmentation procedure.

Finally, the segmentation algorithms were evaluated by using fresh and FFPE heterogeneous samples with 2 or more regions [1–3]. This approach differs from alternative studies that employ homogeneous cancer vs. non-cancer samples for their evaluation procedures [7]. Even though increasing the number of regions could potentially increase the model complexity, chapter 4 implements an RFF approximation technique to reduce the overall computational complexity while maintaining the non-linear data model definition.

5.2. Future Work

This section introduces tentative guidelines for the future direction of this research topic.

Considering the performance gain obtained through the implementation of supervised learning algorithms in chapter 4, alternative supervised algorithms with deep learning could potentially further increase the region detection rates in freshly excised human samples. In particular, CNN has achieved promising results for THz image segmentation in [8–10] due to the integration of the spatial information per pixel in its model definition. Among different CNN techniques, U-Net represents a suitable approach for our application since there is a limited amount of samples for both training and testing procedures. To efficiently utilize the breast cancer samples we have currently available, U-Net proposes the implementation of data augmentation techniques to reduce the size of the training dataset without impacting the overall detection accuracy rate [11]. It is important to consider that the data we have available at the moment may still not be enough for the correct training stage of this procedure, but it would be interesting to consider the usage of multiple tissue models together as a single dataset for training purposes. For instance, the combination of rat and human tissue are likely compatible due to the similarity among the breast cancer tissue that has been collected from these 2 models [12].

To further improve the accuracy of the ground truth definition, it can be beneficial to incorporate x-rays analysis for the definition of the region label per pixel. Even though the pathology analysis of the sample is considered the gold standard of cancer detection, it introduces uncertainty for the ground truth definition of the fresh samples due to the dehydration of the tissue. Considering that x-rays are collected while the tissue is fresh, the information collected from this analysis could be directly compared with the THz image due to the absence of tissue deformation. However, the reliability of x-rays techniques for the detection of cancer is still a topic of research and cannot be considered as ground truth information by itself [13]. Hence, future evaluation procedures within THz imaging of fresh samples should consider the combination of the information collected from both x-rays and histopathology. Similarly to the image morphing procedure introduced in [14],

the control points employed in the algorithm could be determined by analyzing the location of the regions in the x-rays results. Overall, this approach could reduce the manual labor of the selection of the points and increase the reliability of the label assignment by considering points inside the contour of the sample while minimizing the introduction of bias.

References

[1] T. Bowman, T. Chavez, K. Khan, J. Wu, A. Chakraborty, N. Rajaram, K. Bailey, and M. El-Shenawee, "Pulsed terahertz imaging of breast cancer in freshly excised murine tumors," *Journal of Biomedical Optics*, vol. 23, no. 2, p. 026004, 2018. doi: 10.1117/1.JBO.23.2.026004.

[2] T. Chavez, N. Vohra, J. Wu, K. Bailey, and M. El-Shenawee, "Breast cancer detection with low-dimension ordered orthogonal projection in terahertz imaging," *IEEE Transactions on Terahertz Science and Technology*, pp. 1–1, 2019. doi: 10.1109/TTHZ.2019.2962116

[3] T. Chavez, N. Vohra, J. Wu, K. Bailey, and M. El-Shenawee, "Spatial image segmentation for breast cancer detection in terahertz imaging," in *2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting*, 2020, pp. 1157–1158. doi: 10.1109/IEEECONF35879.2020.9330445

[4] T. Chavez, N. Vohra, K. Bailey, M. El-Shenawee, and J. Wu, "Supervised Bayesian learning for breast cancer detection in terahertz imaging," submitted for publication.

[5] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, March 1982. doi: 10.1109/TIT.1982.1056489

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep 1995. doi: 10.1007/BF00994018. [Online]. Available: https://doi.org/10.1007/BF00994018

[7] W. Liu, R. Zhang, Y. Ling, H. Tang, R. She, G. Wei, X. Gong, and Y. Lu, "Automatic recognition of breast invasive ductal carcinoma based on terahertz spectroscopy with wavelet packet transform and machine learning," *Biomed. Opt. Express*, vol. 11,

no. 2, pp. 971–981, Feb 2020. doi: 10.1364/BOE.381623. [Online]. Available: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-11-2-971

[8] Y. Shen, Y. Yin, B. Li, C. Zhao, and G. Li, "Detection of impurities in wheat using terahertz spectral imaging and convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 181, p. 105931, 2021. doi: https://doi.org/10.1016/j.compag.2020.105931. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169920331367

[9] M. Kovbasa, A. Golenkov, and F. Sizov, "Neural network application to the postal terahertz scanner for automated detection of concealed items," in *2020 IEEE Ukrainian Microwave Week (UkrMW)*, 2020, pp. 870–873. doi: 10.1109/UkrMW49653.2020.9252706

[10] H. Feng, D. An, H. Tu, W. Bu, W. Wang, Y. Zhang, H. Zhang, X. Meng, W. Wei, B. Gao, and S. Wu, "A passive video-rate terahertz human body imager with real-time calibration for security applications," *Applied Physics B: Lasers and Optics*, vol. 126, no. 8, p. 143, Aug. 2020. doi: 10.1007/s00340-020-07496-3

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015. ISBN 978-3-319-24574-4 pp. 234–241.

[12] N. Vohra, T. Chavez, J. R. Troncoso, N. Rajaram, J. Wu, P. N. Coan, T. A. Jackson, K. Bailey, and 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*, vol. 8, no. 2, pp. 1 – 17, 2021. doi: 10.1117/1.JMI.8.2.023504. [Online]. Available: https://doi.org/10.1117/1.JMI.8.2.023504

[13] B. Li, Y. Zhang, W. Wu, G. Du, L. Cai, H. Shi, and S. Chen, "Neovascularization of hepatocellular carcinoma in a nude mouse orthotopic liver cancer model: a morphological study using x-ray in-line phase-contrast imaging," *BMC cancer*, vol. 17, no. 1, pp. 1–11, 2017.

[14] T. Chavez, T. Bowman, J. Wu, K. Bailey, and M. El-Shenawee, "Assessment of terahertz imaging for excised breast cancer tumors with image morphing," *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 39, no. 12, pp. 1283–1302, Dec 2018. doi: 10.1007/s10762-018-0529-8