Supervised Statistical Learning for Cancer Detection in Dehydrated Excised Tissue with Terahertz Imaging

Tanny Chavez, Nagma Vohra, Jingxian Wu,
Narasimhan Rajaram, Magda El-Shenawee
University of Arkansas
Fayetteville, AR 72701, USA
tachavez@email.uark.edu, nvohra@email.uark.edu,
wuj@uark.edu, nrajaram@uark.edu, magda@uark.edu

Keith Bailey
University of Illinois at Urbana-Champaign
Urbana, IL 61802, USA
kbailey1@illinois.edu

***Abstract*—This paper proposes a new supervised image segmentation algorithm for the detection of breast cancer using terahertz (THz) imaging. Even though unsupervised learning algorithms have achieved promising results in THz image segmentation, reliable segmentation of tissues with three or more regions, such as cancer, fat and muscle, still remains a major challenge. We propose to tackle this challenge by developing a supervised statistical learning method based on multi-class Bayesian ordinal probit regression. The proposed algorithm utilizes a latent variable for the categorical classification of each pixel within the image. The model parameters are estimated through a Markov chain Monte Carlo (MCMC) process during the training phase. Experimental results in murine formalin-fixed paraffin-embedded (FFPE) breast cancer samples demonstrated that the proposed supervised model outperforms alternative unsupervised methods.**

I. Introduction

Terahertz (THz) imaging has proven to be an effective technique for breast cancer margin detection in breast-conserving surgery [1]–[3]. As shown in our previous work [1], applying unsupervised statistical learning on THz images of breast cancer samples can achieve a $\sim$70% area under the receiver operating characteristic (ROC) curves. However, reaching a 90% ROC area still remains a significant challenge for unsupervised learning techniques, especially for tissues with three or more regions, such as cancer, fat and muscle. An assortment of studies have analyzed different image segmentation techniques based on supervised learning methods, such as support vector machine (SVM) [4], [5], K-nearest neighbors (KNN) [4], [5], random forest (RF) [5], [6], and neural networks [6], [7].

We propose to develop a supervised statistical learning method for the detection of breast cancer in THz imaging. Specially, the supervised statistical learning is developed by using a Bayesian ordinal probit regression model. Unlike common binary regression models, this algorithm incorporates a latent variable that enables the categorical partition of region labels for each pixel [8]. The regression parameters of the model are iteratively learned by using Markov-chain Monte Carlo (MCMC) during a training phase with labeled data. Since the region labels per pixel are known under a supervised context, the learning process is not constrained by the parameters' prior distributions as in the unsupervised learning approach used in our previous works.

For this study, the number of parameters to be learned in the Bayesian ordinal probit model is much less than deep learning approaches such as convolutional neural network (CNN), thus the proposed method requires much less training samples compared to approaches based on deep learning.

II. Method

This paper utilizes xenograft murine breast cancer tumors for the evaluation of the proposed segmentation algorithm. Details on the collection and imaging processes of these samples can be found in [2]. Once the THz image is collected, the high-dimensional waveform per pixel goes through a low-dimensional ordered orthogonal projection (LOOP) process, which projects these waveforms into a lower-dimensional subspace while minimizing the loss of information [1]. Subsequently, the outputs of the LOOP algorithm are processed by applying a categorical probit regression model [8], which is used to perform image segmentation of the sample of interest.

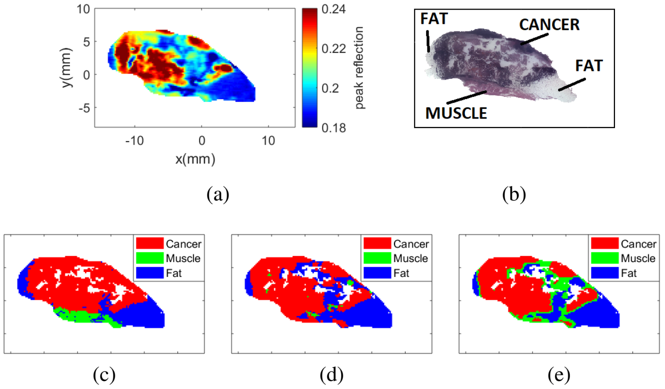
Consider $\mathbf{x} \in \mathcal{R}^L$ to be the lower-dimensional representation of the n-th pixel in the THz image. Given the value of a latent variable, $z_n \in \mathcal{R}$, and a set of thresholds $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \ldots, \alpha_K\}$, where K represents the total number of regions in the sample, the region label of each pixel is assigned according to the range where the latent variable lies among $\boldsymbol{\alpha}$, e.g. the n-th pixel belongs to the k-th region if $\alpha_{k-1} < z_n < \alpha_k$. Based on the assumption that the latent variable and the regression coefficients are linearly correlated, the model definition can be represented as:

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

where $\mathbf{w}_n = [1, \mathbf{x}_n]$, the column vector $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_L] \in \mathcal{R}^{L+1}$ contains the regression coefficients, and σ^2 represents the variance. The corresponding model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ are obtained during the training phase by using a MCMC approach with labeled data. The final category assignment, y_n, is performed as follows:

$$P(y_n = k) = \Phi\left(\alpha_k - \boldsymbol{\beta}^T \mathbf{w}_n, \sigma^2\right) - \Phi\left(\alpha_{k-1} - \boldsymbol{\beta}^T \mathbf{w}_n, \sigma^2\right) \tag{2}$$

This work was funded by the National Institutes of Health under Award No. R15CA208798. It was also funded in part by the National Science Foundation under Award Nos. 1408007 and 1711087.



(a) (b)

(c) (d) (e)

Fig. 1: Mouse 9B FFPE. (a) THz image. (b) Pathology results. (c) Morphed pathology. (d) 2D unsupervised MCMC model. (e) 2D supervised regression model.

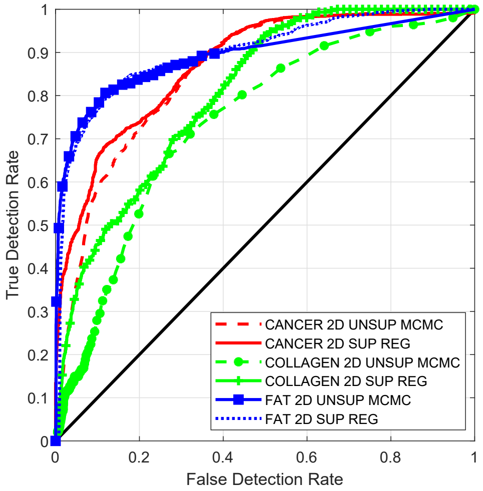


Fig. 2: ROC curves for Mouse 9B FFPE.

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

III. Results

The results presented in this paper correspond to xenograft FFPE sample Mouse 9B [2], which contains cancer, muscle, and fat. During the training stage, the model parameters were estimated by using 3 FFPE murine samples with the same regions, except for one that contained fibro instead of muscle. Fig. 1a represents the THz image collected by using the peak amplitude of the reflected waveform per pixel in the time domain. Fig. 1b corresponds to the pathology analysis of the sample. Fig. 1c represents the morphed pathology mask, which specifies the pixel-by-pixel ground truth [3]. Fig. 1d shows the classification results obtained through the 2-dimensional (2D) unsupervised MCMC approach described in [1]. Fig. 1e presents the classification results obtained by the proposed supervised regression model. As shown in Figs. 1d and 1e, we can observe that the supervised method detects the muscle region better than its unsupervised counterpart.

To quantify the performance of these segmentation algorithms, Fig. 2 presents the ROC curves for mouse 9B FFPE, which shows that the supervised regression model performs better than the unsupervised algorithm for cancer and muscle detection. In addition, Table I presents the areas under the ROC curves for both algorithms. In this table, we can highlight that the supervised regression model obtains areas of 81-90% for all the regions, while the unsupervised model achieved 74-89%.

TABLE I: Areas under the ROC curves for Mouse 9B FFPE.

Method	Cancer	Collagen	Fat
2D unsupervised MCMC	0.8543	0.7398	0.8937
Supervised regression	0.8729	0.8067	0.9007

IV. Conclusion

This paper presents a Bayesian ordinal probit regression algorithm for the detection of breast cancer in THz imaging. The proposed algorithm incorporates a latent variable for the multiclass regression classification of the input data, which requires much less training samples compared to deep learning based approaches. Experimental results demonstrated that the proposed algorithm improves the overall region segmentation in THz images of FFPE murine samples.

Acknowledgment

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

References

[1] T. Chavez, N. Vohra, J. Wu, K. Bailey, and M. El-Shenawee, “Breast cancer detection with low-dimensional ordered orthogonal projection in terahertz imaging,” *IEEE Transactions on Terahertz Science and Technology*, vol. 10, no. 2, pp. 176–189, March 2020. doi: 10.1109/TTHZ.2019.2962116.

[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. 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

[4] 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

[5] 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

[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] 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

[8] 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. doi: 10.1080/01621459.1993.10476321. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476321