Classification and statistical analysis of skin cancer terahertz spectra

M. J. Berryman, T. J. Rainsford and D. Abbott

Centre for Biomedical Engineering and School of Electrical and Electronic Engineering,

The University of Adelaide, SA 5005, Australia

*mattjb@eleceng.adelaide.edu.au

Abstract - **We are developing computational, statistical pattern recognition tools that allow for the separation of various THz spectral data obtained from normal, pre-cancerous and cancerous tissue.**

I. INTRODUCTION

THz is a promising modality for studying skin cancer: A number of international studies have shown that THz spectroscopic and imaging techniques can be used to differentiate between normal and cancerous tissue [1]. Studies using PET and MRI [2] have shown that tumours have increased water content. As water has some prominent resonances in the THz band, imaging in this region is very sensitive to water concentration and increased image contrast has been observed in THz imaging of BCC [3]. THz can also excite many additional energy transitions that correspond to rotational, vibrational and translational modes in complex organic molecules. Studies using near infrared spectroscopy show that pre-malignant and malignant tissue can be differentiated with an accuracy of 71% and that it is possible to differentiate benign from malignant tissue with an accuracy of 85% [4]. Such studies indicate that a continuum of biochemical or structural changes occur in skin cells that are signatures for different cancer types in various stages, from benign to pre-malignant to malignant. THz spectroscopy and imaging has a number of advantages over IR, in particular significantly reduced scattering. Further, THz provides information about the low frequency modes of a molecule and is therefore ideally suited to detect cancer-induced biochemical and structural changes.

II. METHODS USED

Sophisticated computational pattern recognition and related multivariate statistical tools are needed to analyse and classify THz spectral data. Pattern recognition (PR) methods can be divided into two categories: unsupervised and supervised. Unsupervised techniques are used to establish whether any intrinsic clustering exists within a sample group. As it is not necessary to include information pertaining to the class of sample prior to the analysis, the outcome of unsupervised clustering analysis has no sources of human bias [5]. Examples of unsupervised methods include principal components analysis and clustering methods such as hierarchical clustering analysis and k-means clustering [6]. In the supervised PR approach, class identity is given to a training set of sample data in order to optimise the separation between the number of sample classes (for example diseased and normal tissue). Then a second independent 'test' set is used to validate the robustness of any class predictions made using the training set. Examples of supervised PR methods include soft independent modeling of class analogy (SIMCA), partial least squares methods, k-nearest neighbour analysis [7], and support vector machines [8]. Supervised and unsupervised classification schemes are the standard way to differentiate data into discrete and disjoint groups, which is the kind of separation we seek for THz spectra data.

REFERENCES

[1] R. M. Woodward, V. P. Wallace, R. J. Pye, B. E. Cole, D. D. Arnone, E. H. Linfield, and M. Pepper, "Terahertz pulse imaging of ex vivo basal cell carcinoma," *J. Invest. Dermatol.*, vol. 120, pp. 72-78, 2003.

[2] E. K. Rofstad, E. Steinsland, O. Kaalhus, Y. B. Chang, B. Hovik, and H. Lyng, "Magnetic resonance imaging of human melanoma xenografts in vivo: proton spin lattice, spin-spin relaxation times verses fractional water content and fraction of necrotic tissue," *Int. J. Radiat. Biol.*, vol. 65, 1994.

[3] R. M. Woodward, B. E. Cole, and V. P. Wallace, "THz pulse imaging of in vitro basal cell carcinoma samples," *TOPS*, vol. 56, pp. 329-330, 2001.

[4] L. M. McIntosh, M. Jackson, H. M. Mantsch, J. R. Mansfield, A. N. Crowson, and J. W. P. Toole, "Near-infrared spectroscopy for dermatological applications," *Vibrational Spectroscopy*, vol. 28, pp. 53-58, 2002.

[5] H. Antti, E. Holmes, and J. Nicholson, "Multivariate solutions to metabonomic profiling and functional genomics: part 2 – chemometric analysis and future perspectives," 2002.

[6] C. M. Bishop, *Neural networks for pattern recognition.* Oxford, UK: Oxford University Press, 1995.

[7] G. Karypis, E. H. Han, and V. Kumar, "Chameleon: a hierarchical clustering algorithm using dynamic modeling," *IEEE Computer*, vol. 32, pp. 68-75, 1999.

[8] O. Mangasarian, "Generalized support vector machines," in *Advances in Large Margin Classifiers*, A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 2000, pp. 135-146.