1-1-2007

Classification of terahertz data as a tool for the detection of cancer

Matthew J. Berryman
University of Adelaide, mberryma@uow.edu.au

T. J. Rainsford
University of Adelaide

Follow this and additional works at: <https://ro.uow.edu.au/engpapers>

Part of the [Engineering Commons](#)

<https://ro.uow.edu.au/engpapers/4951>

Recommended Citation

Berryman, Matthew J. and Rainsford, T. J.: Classification of terahertz data as a tool for the detection of cancer 2007.
<https://ro.uow.edu.au/engpapers/4951>

Cancer screening using Terahertz imaging technology and advanced pattern recognition

Tamath Rainsford and Matthew J. Berryman

Terahertz spectroscopic and imaging systems when combined with computational pattern recognition and related multivariate statistical tools potentially provide a fast and non-invasive way to detect and diagnose diseases on the surface of the skin such as cancer.

With the advent of new technologies that are both noninvasive and rapid, comprehensive screening and diagnosis for common diseases are likely to become routine in the future. Terahertz (THz) spectroscopy and imaging are one such technology. THz is radiation in the range 3×10^{11} Hz (submillimeter end) to 3×10^{12} Hz (far-infrared end) and can penetrate a wide variety of non-conducting materials. As it is non-ionizing and, is able to penetrate several millimeters of living tissue and reflect back, it is a potentially useful modality for investigating diseases on the surface of the skin such as skin cancer. Further, unlike visual inspection, THz images contain spectral information on both amplitude and phase in each image pixel. By linking spectral with spatial information, THz images provide a more detailed spatial representation of the bio-molecular content in a sample than a conventional image, taken either visually or with X-rays.

The difficulties with using THz as a diagnostic tool lie not so much in obtaining spectroscopic data and images but in how to obtain useful information from them. One looks for changes in the THz pulse that are caused by the sample. However these can be difficult to distinguish from those caused by long term fluctuations in the driving laser source. Apart from this little is known about the interaction of THz with tissue at the molecular level. THz can produce images of samples which are opaque in the visible and near-infrared regions of the spectrum often, showing increased contrast which allows for visual identification of a sample. It has been shown that THz images can be used to differentiate between normal and cancerous tissue¹⁻³. Studies using PET and MRI⁴ 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 BCC¹. However, the visual markers that are used to distinguish different cancers and developmental stages can be subtle. Vast information may lie in the THz spectroscopic responses of tissue samples as many biomolecules have resonant peaks due to phonon and intermolecular vibrations in the THz range⁵ and sophisticated computational pattern recognition and related multivariate statistical tools are needed to analyse and classify these data.

Pattern classification is a technique for assigning raw data to one of several classes. Typically raw data is preprocessed before being provided to a classifier. This is usually for de-noising and normalizing. The actual techniques used to do this are specific to the type of data to be processed. The first step in classification is feature extraction and it is this step that leads to either a trivial problem or to something potentially impossible to solve. The features are provided to a classifier which is trained with knowledge about the potential classes. There are a large number of features that could be chosen and the both the types of features and the number of features used can lead to very different results. It is also possible to use unsupervised techniques where the training step is omitted. There are many different classification algorithms which differ greatly in terms of their complexity and sophistication.

Getting more technical

We have explored a highly sophisticated technique for classification, Support Vector Machines (SVM), on a simple data set. Our data set consisted of THz spectral data collected from osteoblast (normal bone tissues) cells and osteosarcoma (a type of bone cancer) cells using a THz-TDS imaging system. Details can be found in ⁶. We preprocessed the raw data using Finite Impulse Response (FIR) filter coefficients. FIR filters⁷ were used here as they capture the system in question: subtracting the effect of the cell media and air and moisture surrounding the sample; they introduce a linear phase-shift to the pulses, which preserves the frequency response and the features of the time-domain response (other than a consistent time, or phase shift); and they

Continued on next page

separate out subtle differences between the normal and cancerous cells, which the support vector machines can then use in effectively separating out the data.

Support Vector Machines (SVM) are one form of machine learning^{8,9} that can be used to classify a set of data into two or more categories. A learned model of the data is created from a training set consisting of data vectors x (one for each sample) with a known label (or category) y . The model is then tested on a different test set over which the model predicts a set of labels, which are then compared with known labels. Of interest is the success rate of prediction of the model. The goal of SVM learning is to find the optimal boundary that separates the clusters of vectors in such a way that features of one group of samples are on one side of the boundary (or hyper-plane) and features of a different group of data are on the other side. The vectors near the boundary are known as the support vectors and it is these vectors that are used to construct the boundary, which is then formulated as a classifier function. In most cases, however, the vectors are not easily separated by a linear plane. Instead, they may be better represented by a non-linear surface (in general, these are n -dimensional problems with a high n). The beauty of SVM is that such complex boundaries can still be constructed by using the same algorithm but with a different kernel function. It can take on many mapping functions with some of the most common types being the linear, polynomial, radical basis function (RBF) and sigmoid types. Therefore, the kernel function allows SVM to perform separations even in the presence of complex boundaries.

We ran the FIR filtering on each pair of sample and reference pulses, examples of these are in Figure 1, then we fed the FIR filter coefficients into the support vector machine for training and testing of the support vector machine, for a total of 1000 iterations, randomizing the selection of training and test sets (from a total possibility of $\binom{48}{30}^2$ different selections). This provides a good measure of the accuracy and its standard deviation, and also compensates somewhat for the low number of samples available for both training and testing. The SVM using the FIR filter coefficients as input were able to distinguish the normal and cancerous cells with an accuracy of 70.2% which is reasonable given the extremely small sample size. While this not a reliable figure, it is substantially better than visual screening but is much lower than histochemical diagnosis which is of the order 85%. These results show that, in principle, THz spectral data can be collected in a noninvasive way that can then be automatically analysed using a classifier to say whether or not a sample is cancerous. The actual features used to do this are not necessarily physical. Visual observation, which is the current method of screening, is based on physical features. A larger data set would likely lead to a more accurate classifier which could be used to answer more specific questions such as, to the type of cancer or

(a) Time domain THz plot of a normal bone cells sample, reference pulse, and FIR filtered reference pulse.

(b) Time domain THz plot of a cancerous bone cells sample, reference pulse, and FIR filtered reference pulse.

Figure 1. Plots of the normal and cancer cell data. The filtered reference pulse should look like a time-delayed version of the sample pulse, and it does but with noticeable differences. Compared with the normal cell data, there is more of a drop in amplitude, and more of a phase difference between the reference and sample pulses, most likely due to more water in the cancer cells compared with the normal cells, in line with Woodward et al.^{1,2,16}.

developmental stage of a cancer.

Conclusions

Cancer is a difficult disease to screen and diagnose, and our lab has made significant progress in this area by combining developments in Terahertz radiation technology with sophisticated computer algorithms, that could screen cancer at a much better success rate than the current screening processes. Diagnosis is another area where Terahertz technology can be used to distinguish cancerous from normal or pre-cancerous tissue.

Other members of our group have been working on methods for removing noise from Terahertz data that improves classification^{10,11} and improvements to the underlying technology¹², in addition to our labs other research in areas such

Continued on next page

as neuron (and neuron-inspired) signalling^{13,14} and genetics¹⁵.

Author Information

Tamath Rainsford, and Matthew Berryman, and Centre for Biomedical Engineering
The University of Adelaide
Adelaide, SA, Australia

Tamath Rainsford is a Research Fellow at the University of Adelaide. Rainsford received a PhD in Mathematical Physics from the University of Adelaide (2000) and has since worked in the biomedical and bioengineering fields.

Matthew Berryman is a Research Scientist at the Australian Defence Science and Technology Organisation, and assists with projects at the Centre for Biomedical Engineering, University of Adelaide, where he is just about to submit his PhD in the field of Complex Systems. He has worked extensively in the area of biological analysis and modelling.

References

1. R. M. Woodward, B. E. Cole, and V. P. Wallace, *THz pulse imaging of in vitro basal cell carcinoma samples*, **TOPS** **56**, pp. 329–330, 2001.
2. R. M. Woodward, B. Cole, V. P. Wallace, R. Pye, D. D. Arnone, E. H. Linfield, and M. Pepper, *Terahertz pulse imaging in reflection geometry of human skin cancer and skin tissue*, **Phys. Med. Biol.** **47** (21), pp. 3853–3863, 2002.
3. E. Pickwell, B. E. Cole, and A. J. Fitzgerald, *In vivo study of human skin using pulsed THz radiation*, **Phys. Med. Biol.** **49** (9), pp. 1595–1607, 2004.
4. 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 versus fractional water content and fraction of necrotic tissue*, **Int. J. Radiat. Biol.** **65** (387–401), 1994.
5. B. M. Fischer, M. Franz, and D. Abbott, *T-ray biosensing: a versatile tool for studying low-frequency intermolecular vibrations* **Biomedical Applications of Micro- and Nanoengineering III** **6416** (1), SPIE, 2006. doi:10.1117/12.695726
6. M. J. Berryman and T. Rainsford, *Classification of terahertz data as a tool for the detection of cancer* **Biomedical Applications of Micro- and Nanoengineering III** **6416** (1), SPIE, 2006. doi:10.1117/12.695693
7. E. C. Ifeachor and B. W. Jervis, **Digital signal processing: a practical approach**, 2nd ed., Prentice Hall, New York, NY, USA, 2002.
8. N. Christianini and S. J. Taylor, **An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods**, Cambridge University Press, 2000.
9. C. Burges, *A tutorial on support vector machines for pattern recognition*, **Data Mining and Knowledge Discovery** **2**, pp. 121–167, 1998.
10. W. Withayachumnankul, B. Ferguson, T. Rainsford, D. Findlay, S. P. Micken, and D. Abbott, *T-ray relevant frequencies for osteosarcoma classification* **Photonics: Design, Technology, and Packaging II** **6038** (1), SPIE, 2005. doi:10.1117/12.637964
11. W. Withayachumnankul, B. Ferguson, T. Rainsford, S. P. Micken, and D. Abbott, *Material parameter extraction for terahertz time-domain spectroscopy using fixed-point iteration* **Photonic Materials, Devices, and Applications** **5840** (1), pp. 221–231, SPIE, 2005. doi:10.1117/12.612946
12. W. Withayachumnankul, B. M. Fischer, S. P. Micken, and D. Abbott, *Retrofittable T-ray antireflection coatings* **Micro- and Nanotechnology: Materials, Processes, Packaging, and Systems III** **6415** (1), SPIE, 2006. doi:10.1117/12.695949
13. M. D. McDonnell, S. Sethuraman, L. B. Kish, and D. Abbott, *Cross-spectral measurement of neural signal transfer* **Noise in Complex Systems and Stochastic Dynamics II** **5471** (1), pp. 550–559, SPIE, 2004. doi:10.1117/12.546926
14. M. D. McDonnell, D. Abbott, and C. E. Pearce, *Neural mechanisms for analog-to-digital conversion* **BioMEMS and Nanotechnology** **5275** (1), pp. 278–286, SPIE, 2004. doi:10.1117/12.523165
15. M. J. Berryman, A. Allison, and D. Abbott, *Mutual information for examining correlations in DNA*, **Fluctuation and Noise Letters** **4** (2), pp. L237–L246, 2004.
16. 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.** **120** (1), pp. 72–78, 2003.