DCT-SVM based multi-classification of mouse skin precancerous stages from autofluorescence and diffuse reflectance spectra

F. Abdat[a], M. Amouroux[a], Y. Guermeur[b] and W. C. P. M. Blondel[a]

[a]Centre de Recherche en Automatique de Nancy (CRAN), UMR 7039, Nancy University-CNRS, 2 avenue de la forêt de haye, 54500 Vandœuvre-lès-Nancy, France
[b]Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), UMR 7503, Nancy University-CNRS, Campus Sciences, BP239, 54506 Vandœuvre-lès-Nancy Cedex, France

ABSTRACT

This paper deals with multi-classification of skin precancerous stages based on bimodal spectroscopy combining AutoFluorescence (AF) and Diffuse Reflectance (DR) measurements. The proposed data processing method is based on Discrete Cosine Transform (DCT) to extract discriminant spectral features and on Support Vector Machine to classify. Results show that DCT gives better results for AF spectra than for DR spectra. This study shows that bimodality and monitoring spectral resolution together allow an increase in diagnostic accuracy. The choice of an adequate spectral resolution always implies an increase in diagnostic accuracy. This accuracy can get as high as 79.0% when combining different distances between collecting and exciting optical fibers.

Keywords: autofluorescence spectroscopy, diffuse reflectance spectroscopy, discrete cosine transform, support vector machine, skin cancer.

1. INTRODUCTION

Precancerous and cancerous progressions in human tissues are characterized by morphological and functional modifications of the biological tissue's constituents (cells, extracellular matrix, and vasculature). Different morphofunctional conditions of the skin related to pathologic alterations such as cancer development can be characterized *in vivo* non invasively using either AutoFluorescence Spectroscopy (AFS), Diffuse Reflectance Spectroscopy (DRS) or both.[20] As shown in previous works,[4,5] combining both modalities (DRS and AFS) provides a more accurate diagnosis.

In order to extract diagnostic useful information, dedicated spectral data analysis algorithms need to be developed including preprocessing, data reduction, identification as well as selection of discriminant spectral features, and classification. Multivariate statistical techniques such as principal component analysis (PCA),[9,14] Linear Discriminant Analysis (LDA),[17] and Artificial Neural Networks (ANNs)[3] have been previously used in efficient diagnostic algorithms for *in vivo* cancerous tissue classification based on optical spectroscopy data. For example, a high diagnostic accuracy (>90.0%) can be achieved for identifying Raman spectra of cancerous from normal tissues in the colon, employing PCA-LDA techniques.[17] Other powerful multivariate models are Support Vector Machines (SVMs), which have proved to provide efficient solutions for the classification of spectral data for tissue diagnosis. For instance, Lin et al[13] used a SVM to differentiate *in vivo* autofluorescence spectra of NasoPharyngeal Carcinoma (NPC) from normal tissues with a diagnostic accuracy higher than that of PCA-LDA. Palmer et al[18] used a linear SVM for identifying AutoFluorescence and Diffuse Reflectance spectra of breast cancer tissues *in vitro*. Majumder et al[15] applied linear and non-linear SVMs to fluorescence spectra for distinguishing between malignant and normal tissues in the oral cavity.

In previously published works,[2,7] our team showed that spectral features could be extracted from *a priori* shape knowledge of spectra (namely slopes, intensities, areas under the curve) and that PCA could be efficiently

Send correspondence to: abdat_faiza@hotmail.com, marine_amouroux@yahoo.fr, yann.guermeur@loria.fr, Walter.Blondel@ensem.inpl-nancy.fr

used to reduce the number of these extracted features. Finally, several classification algorithms were compared for discriminating between pairs of histological classes: k-Nearest Neighbors (kNN), LDA, and SVM. Most of the studies to date on the application of optical spectroscopy to skin focused on a local analysis with *a priori* knowledge on curve-shape features that allow the distinction between histological classes.

The contribution of the present study is focused on the global analysis of the raw AF and the DR spectra without any prior knowledge on a specific wavelength for a multi-class distinction of skin precancerous stages. This extend our previous work by discriminating between the different histological classes using SVM based multi-class classification. We also propose to base both feature extraction and feature reduction on Discrete Cosine Transform (DCT). The aim is to replace the initial set of features by a smaller one (where the new features are linear combinations of the initial ones). The strong capability of DCT to compress energy makes it a good candidate for pattern recognition applications. It allows more flexibility in the choice of the new variables which is a potential advantage.

The raw data set consists of 252 samples which are spots of mouse skin either sham-irradiated (control group) or UV-irradiated once then twice a week for up to 20 weeks. Each spot is initially described by 24 spectra acquired *in vivo*:7 AF spectra (7 excitation wavelengths: 360, 368, 390, 400, 410, 420, and 430 nm) and 1 DR spectrum (UV-visible spectral range) at 3 different distances between collecting and exciting optical fibers (Collection to Excitation Fiber Separation CEFS): D1=271 μm, D2=536 μm, and D3=834 μm. Four reference classes were defined based on the histopathological analysis of the skin samples: Healthy (H), Compensatory Hyperplasia (CH), Atypical Hyperplasia (AH), and Dysplasia (D). After *in vivo* spectral data acquisition, the 252 mouse skin spots were biopsied and classified by an anatomopathologist physician into the 4 histological classes: 84 were considered as H, 47 were classified into CH class, 64 into AH class, and 57 into D class.[2]

2. DATA PROCESSING

2.1 Spectral feature extraction using Discrete Cosine Transform

The DCT has been used in many practical applications, especially in signal compression (JPEG). For k in $\{1,\ldots,N\}$, the *kth* DCT coefficient $x_c(k)$ of a spectrum $(x(n))_{1\leqslant n\leqslant N}$ is defined as[12,19]:

$$x_c(k) = w(k)\sum_{n=1}^{N} x(n)\cos\left[\frac{\pi(2n-1)(k-1)}{2N}\right] \tag{1}$$

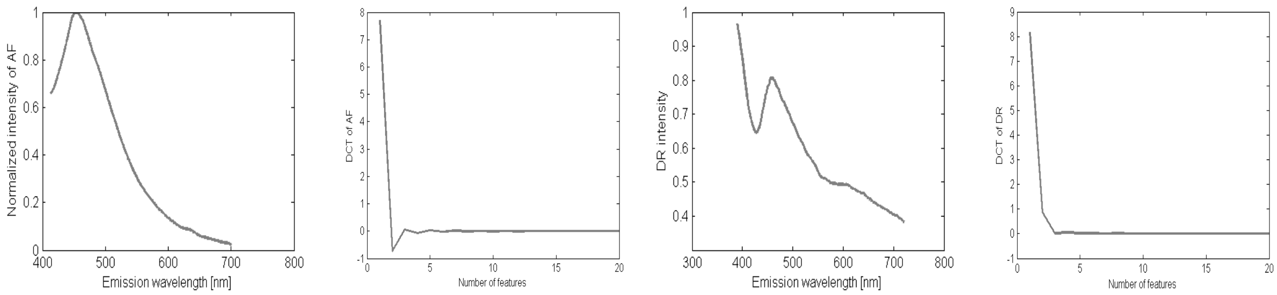
where $w(k) = \begin{cases} \frac{1}{\sqrt{N}} & if \quad k=1 \\ \sqrt{\frac{2}{N}} & 2\leqslant k\leqslant N \quad otherwise \end{cases}$

It can be shown that $x_c(k)$ is a bandpass filter with a center frequency at $(2k+1)/2N$ when the sampling frequency is normalized to 1. Hence, the magnitude of the output $x_c(k)$ for small values of k is generally larger.

In other words, one can restrict the expansion to the terms of low indices without a significant impact on the energy of the signal.[1,11] As shown in Figures 1-b, and 1-d, lots of frequency components of AF spectra (see example in Figure 1-a) or of DR spectra (see example in Figure 1-c) are concentrated at the lowest frequencies. Most of the visually significant information is concentrated in just a few first DCT coefficients, which we selected as features.

2.2 Support Vector Machine based classification

In order to classify automatically the 252 tissue samples within the 4 classes (H, CH, AH, D) based on their spectral features previously extracted, a multi-category supervised classification was performed with reference to the "gold standard" results of histological analysis. The accuracy of the classification is characterized by means of the recognition rate τ, which is studied as a function of the number of features.



a b c d

Figure 1. Examples of an emission autofluorescence spectrum for an excitation wavelength of 360 nm (a) and a diffuse reflectance spectrum (c), both acquired at inter-fiber distance D1. (b) and (d) show the 20 first DCT coefficients calculated for (a) and (c) respectively.

Several methods have been used for multi-class classification such as Artificial Neural Networks, kNN, and SVMs. The latter have rapidly gained in popularity due to their theoretical merits, computational simplicity, and excellent performance in real-world applications.[23] The first studies dealing with the use of SVMs for multi-category classification report results obtained with decomposition methods[22,25] involving Vapnik's machine. Multi-class support vector machines (M-SVMs) were only introduced three years later.[27] We have chosen to implement the One-Versus-All (OVA) decomposition scheme[21] with bi-class SVMs as base classifiers. All machines use a Gaussian kernel given by:

$$\kappa\left(y, y'\right) = \exp\left(-\mu \left\|y - y'\right\|^2\right) \tag{2}$$

where y and y' are two data points, and μ is the bandwidth coefficient. These choices were motivated by the results of initial experiments involving different multi-class SVMs,[8] different decomposition methods, and different kernels (data not shown). The experimental protocol is systematically a 5-fold cross-validation.

3. RESULTS AND DISCUSSION

3.1 Determining an appropriate number of DCT coefficients

Selecting the discriminant features (extracting relevant information) is the first key step of our work. If the features extracted are carefully chosen it is expected that they will extract the relevant information for the task of interest. This allows to prevent the curse of dimensionality, avoiding classification errors produced by domination of irrelevant features.[16] In order to choose an appropriate number of DCT coefficients, the recognition rate was computed as a function of the number of DCT coefficients retained. Experimental results are summarized in Figures 2-a, 2-b, and 2-c for AF spectra alone, the DR spectra alone, and AF and the DR spectra together (bimodality). For AF spectra, a slight increase in recognition accuracy was observed as we go to higher numbers of coefficients. For the DR spectra, the prediction accuracy increases with the number of coefficients until it reaches a plateau for N_{DR}=147 coefficients. From this value of N_{DR}, no change in recognition accuracy was observed.

For bimodality configuration, a slight decrease in recognition accuracy was observed as we go to higher numbers of coefficients. In other words, more DCT coefficients does not necessarily mean better recognition rate because high frequency components are related to unstable activity features such as noise. According to Figures 2-a, and 2-b, the best performance of our system is obtained when N_{AF}=56 for AF and when N_{DR}=147 for DR. We calculated the recognition rate with 56 coefficients for each spectrum multiplied by 7, the number of excitation wavelengths, and by 3, the number of CEFS. With these 1176 features, we obtained a rate of $\simeq$78.0% for the combination D1D2D3, which is close to that found with D1 alone. This result can be explained by the fact that the number of features is much larger than the cardinality of the data set (252 samples) used for learning. Thus, the curse of dimensionality could be a problem impacting our classifier. To overcome this problem, we

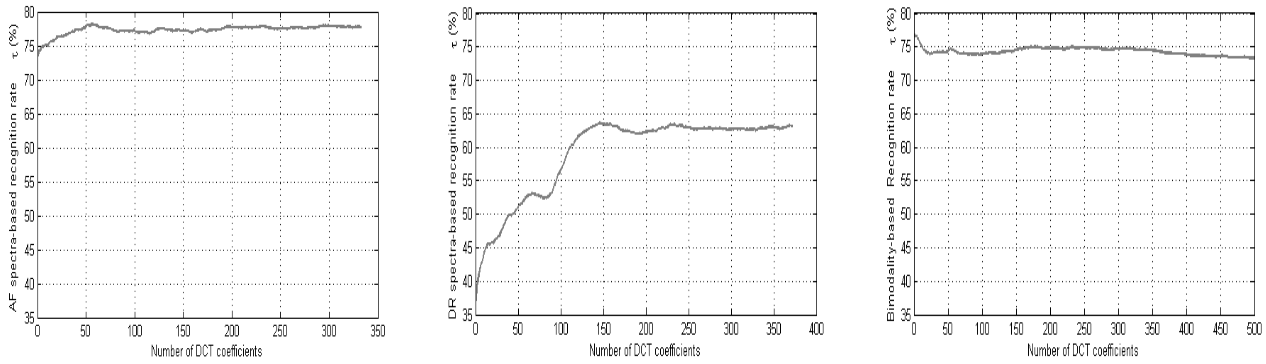


Figure 2. Recognition rate τ (%) as a function of the number of DCT coefficients retained. From left to right: AF spectra alone acquired at D1, the DR spectrum alone acquired at D1, and AF and the DR spectra together (bimodality).

worked with $N_{AF} = 2$ per AF excitation (i.e., 2x7=14 in total). For DR, we kept working with $N_{DR} = 147$ (see Figure 2-b).

3.2 Spatial resolution effect on the classifier performance

Figure 3 shows results obtained from the different sets of characteristics for each mono- and bimodal excitation. Each bar displays the SVM results for each of the 7 possible combinations of CEFS tested: D1 alone, D2 alone, D3 alone, D1D2, D1D3, D2D3, and D1D2D3 with the aforementioned sets of features (N_{AF} and N_{DR}). Results obtained with D1 alone are globally better than those obtained with D2 alone or with D3 alone. The diagnostic accuracy obtained with D1 is the best for each mono-modal excitation (74.5% for AF and 63.6% for the DR), as well as for bimodal approach (76.7%).

The choice of an adequate spectral resolution always implies an increase in diagnostic accuracy. For each modality, an increase in discrimination accuracy of the 4 histological classes was obtained using an extended set of characteristics from multiple distances D1D3, D1D2, and D1D2D3. The combination D2D3 gives the lowest rate for each modality compared to D1D3, D1D2, and D1D2D3.

An increase in discrimination accuracy was obtained using a more extended set of characteristics D1D2D3 for bimodal approach compared to each mono-modal excitation (AF or DR). The best result of our bimodal system, 79.0%, was obtained with the combination D1D3. The performance of our system depends on two factors. The first one is the choice of the CEFS combination for each modality. The second one is the selection of the coefficients for DCT.

Overall, the DCT gives better results for the AF modality than for the DR one. For this reason, we intend to turn to a hybrid method, where DCT would be applied to AF spectra and another method to the DR spectrum. The results obtained by our team in the previously published work[7] are the following: the sensibility value Se and the specificity value Sp reached 100% when discriminating CH vs. others; $Sp \approx 100.0\%$ and $Se > 95.0\%$ when discriminating Healthy vs. AH or D; $Sp \approx 74.0\%$ and $Se \approx 63.0\%$ for AH vs. D.

3.3 Excitation wavelength effect on the classifier performance

Since DCT method gave good results with AF modality alone, preserving useful information and allowing a global analysis of the spectrum, the recognition rate for each excitation wavelength using this feature extraction method was calculated. Figures 4-a, 4-b, and 4-c represent the variation of the recognition rate for the different excitation wavelengths used with D1, D2, and D3 respectively. In general, all excitation wavelengths (except that of 420 nm) gave similar results, i.e., the higher the number of coefficients, the higher the recognition rate. However, the 420 nm excitation wavelength leads to stable performance regardless of the inter-fiber distance. These results

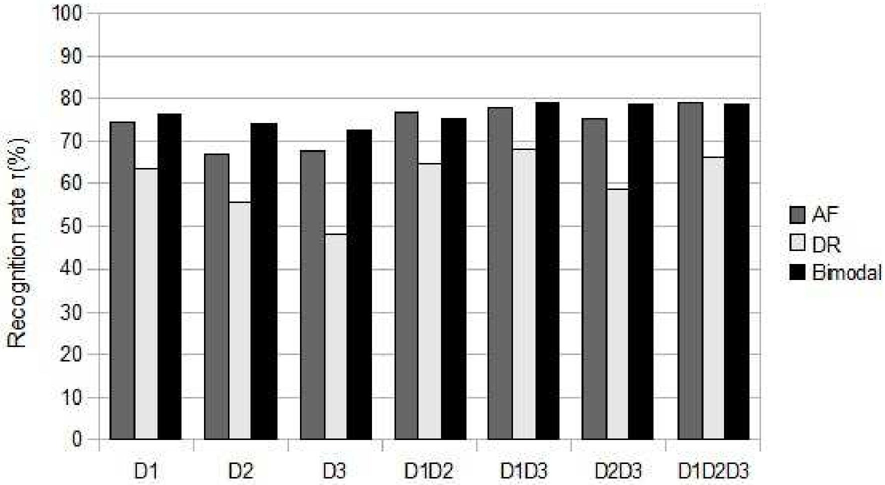
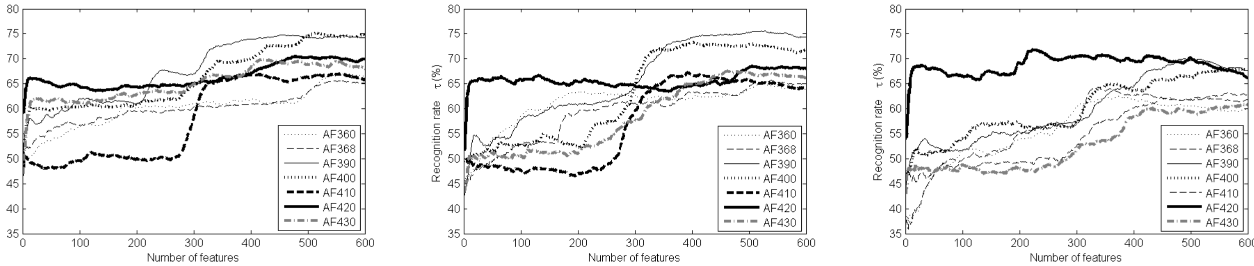


Figure 3. Recognition rate for the different combinations of modalities and CEFS.



(a) (b) (c)

Figure 4. Recognition rate according to the number of DCT coefficients for several excitation wavelengths of AF modality with (a) D1 (b) D2 (c) D3.

confirm those obtained by Van Staveren[24] who chose this single excitation wavelength to distinguish between different grades of oral tissues. For the latter study, the choice of 420 nm excitation wavelength was guided by previous investigations with excitation wavelengths in the range of 400-450 nm.[6, 10] At this wavelength, several endogenous fluorophores are excited, viz. porphyrins, lipo-pigments, and flavins, which have a fluorescence emission in the range of 450-650 nm.[26]

According to Figures 4-a, 4-b, and 4-c the best rates with 420 nm excitation (AF420) are around 65.0%, hence the importance of multi excitation that allows to achieve a rate that exceeds 74.0%.

3.4 Confusion matrix

Table 1 shows the classification results in the form of the confusion matrix that compares the pathologist's diagnosis (rows) with that of the DCT-SVM-based spectral data classification (columns). The best performance of the latter was in classifying H vs all and CH vs all tissues (100.0% accuracy), and errors were spread among the remaining classes D and HA.

The highest confusion is observed between HA and D, 31.0% of HA are classified as D and 47.0% of D are classified as HA. Both HA and D are malignant classes, so, the confusion between these two classes does not alter the diagnosis, whereas the confusion between malignant and healthy classes has an effect on the diagnosis accuracy. Thus, 19.0% of HA and 11.8% of D are classified as H, which corresponds to incorrect predictions (false-negative), so, $Se = 81.0\%$ for D and $Se = 88.2\%$ for HA, with $Sp = 100.0\%$.

The overall intensity of the spectral curve of the class H is very different from that of the three other, whereas the intensities of D and HA are close to each other. Our data set is characterized by a continuous distribution

		SVM			
		H	CH	HA	D
Histo	H	100.0%	0.0%	0.0%	0.0%
	CH	0.0%	100.0%	0.0%	0.0%
	HA	19.0%	0.0%	50.0%	31.0%
	D	11.8%	0.0%	47.0%	41.2%

Table 1. Confusion matrix calculated with DCT coefficients of AF and the DR spectra acquired at D1

of normal and low grade lesions, which makes the separation between classes difficult.

4. CONCLUSION AND FUTURE WORK

In this paper we have introduced a new bimodal system to recognize skin precancerous stages using autofluorescence and diffuse reflectance spectroscopies. The spectral data processing performed include the extraction of discriminant features, and the classification step. The added value of this paper rests in the fact that it tackles multi-class problems because most of the studies to date on the application of optical spectroscopy to skin cancer diagnosis focused on binary classification problems. The method used for feature extraction is discrete cosine transform. Support Vector Machine was used to classify skin samples into four histopathological classes. Results show that DCT gives better results for AF spectra than for the DR spectrum. Using a bimodal approach and appropriate spectral resolution together allows an increase in diagnostic accuracy. This accuracy can get as high as 79.0% when combining different distances between collecting and exciting optical fibers. Our future work will include the use of other Multi-class SVMs. Moreover, we plan to devise dedicated kernels for our application.

Acknowledgment

This work was performed with financial support from the CNRS under a postdoctoral fellowship N°225412.

REFERENCES

1. N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 23(1):90–93, 1974.
2. M. Amouroux, G. Diaz-Ayil, W.C.P.M. Blondel, G. Bourg-Heckly, A. Leroux, and F. Guillemin. Classification of ultra-violet irradiated mouse skin histological stages by bimodal spectroscopy (multiple excitation autofluorescence and diffuse reflectance). *Journal of Biomedical Optics*, 14(1):14011–14024, 2009.
3. J. Backhausa, R. Muellera, N. Formanskia, N. Szlamaa, H. G. Meerpohlb, M. Eidtb, and P. Bugertc. Diagnosis of breast cancer with infrared spectroscopy from serum samples. *Vibrational Spectroscopy*, 52(2):173–177, 2010.
4. S.K. Chang, Y.N. Mirabal, E.N. Atkinson, D. Cox, A. Malpica, M. Follen, and R. Richards-Kortum. Combined reflectance and fluorescence spectroscopy for *in vivo* detection of cervical pre-cancer. *Journal of Biomedical Optics*, 10(2):024031, 2005.
5. D.C.G. de Veld, M. Skurichina, M.J.H. Witjes, R.P.W. Duin, H.J.C.M. Sterenborg, and J.L.N. Roodenburg. Autofluorescence and diffuse reflectance spectroscopy for oral oncology. *Lasers in Surgery and Medicine*, 36:356–364, 2005.
6. J.K. Dhingra, D.F. Perrault, K. McMillan, E.E. Rebeiz, S. Kabani, R. Manoharan, I. Itzkan, M.S. Feld, and S. M. Shapshay. Early diagnosis of upper aerodigestive tract cancer by autofluorescence. *Archives of Otolaryngology-Head and Neck Surgery*, 122(11):1181–6, 1996.
7. G. Diaz-Ayil, M. Amouroux, W.C.P.M. Blondel, G. Bourg-Heckly, A. Leroux, F. Guillemin, and Y. Granjon. Bimodal spectroscopic evaluation of ultra violet-irradiated mouse skin inflammatory and precancerous stages: instrumentation, spectral feature extraction/selection and classification (k-NN, LDA and SVM). *The European Physical Journal-Applied Physics*, 47(1):12707–12718, 2009.
8. Y. Guermeur. A generic model of multi-class support vector machine. *International Journal of Intelligent Information and Database Systems*, 2011. (accepted).

9. P. K. Gupta, S. K. Majumder, and A. Uppal. Breast cancer diagnosis using N2 laser excited autofluorescence spectroscopy. *Lasers in Surgery and Medicine*, 21(5):417–422, 1997.
10. M.L. Harries, S. Lam, C. Macaulay, J. Qu, and B. Palcic. Diagnostic imaging of the larynx: autofluorescence of laryngeal tumours using the helium-cadmium laser. *Journal of Laryngology and Otology*, 109:108–110, 1995.
11. Z. He and L. Jin. Activity recognition from acceleration data based on discrete consine transform and SVM. *IEEE International Conference on Systems, Man and Cybernetics*, 2009.
12. A.K Jain. *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
13. W.M. Lin, X. Yuan, P. Yuen, W. I. Wei, J. Sham, P. C. Shi, and J. Qu. Classification of *in vivo* autofluorescence spectra using support vector machines. *Journal of Biomedical Optics*, 9(1):180–186, 2004.
14. S. K. Majumder, P. K. Gupta, B. Jain, and A. Uppal. UV excited autofluorescence spectroscopy of human breast tissues for discriminating cancerous tissue from benign tumor and normal tissue. *Lasers in the Life Sciences*, 8:249–264, 1999.
15. S.K. Majumder, N. Ghosh, and P.K. Gupta. Support vector machine for optical diagnosis of cancer. *Journal of Biomedical Optics*, 10(2):024034, 2005.
16. M.O. Mendez, A.M. Bianchi, M. Matteucci, S. Cerutti, and T. Penzel. Sleep apnea screening by autoregressive models from a single ecg lead. *IEEE Transactions on Biomedical Engineering*, 56(12):2838–2850, 2009.
17. A.L.M. Molckovsky, K.S. Wong, M.G. Shim, N.E. Marcon, and B.C. Wilson. Diagnostic potential of near-infrared Raman spectroscopy in the colon: differentiating adenomatous from hyperplastic polyps. *Gastrointestinal Endoscopy*, 57:396–402, 2003.
18. G. M. Palmer, C. Zhu, T. M. Breslin, F. Xu, K. W. Gilchrist, and N. Ramanujam. Comparison of multiexcitation fluorescence and diffuse reflectance spectroscopy for the diagnosis of breast cancer. *IEEE Transactions on Biomedical Engineering*, 50:1233–1242, 2003.
19. W.B. Pennebaker and J.L. Mitchell. *JPEG Still Image Data Compression Standard*, chapter 4. New York, NY: Van Nostrand Reinhold, 1993.
20. L.T. Perelman and V. Backman. *Handbook of optical biomedical diagnostics*, chapter 12, pages 675–724. Tuchin V, ed., Bellingham: SPIE Press, 2002.
21. R. Rifkin and A. Klautau. In defense of One-Vs-All classification. *Journal of Machine Learning Research*, 5:101–141, 2004.
22. B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. *International conference on Knowledge Discovery and Data Mining*, pages 252–257, 1995.
23. B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA, 2002.
24. H. J. van Staveren, R. L. P. van Veen, O. C. Speelman, M. J. H. Witjes, W. M. Star, and J. L. N. Roodenburgb. Classification of clinical autofluorescence spectra of oral leukoplakia using an artificial neural network: a pilot study. *Oral Oncology*, 36(3):286–293, 2000.
25. V.N. Vapnik. The nature of statistical learning theory. *Springer-Verlag, New York*, 1995.
26. G.A. Wagnieres, W.M. Star, and B.C. Wilson. *In vivo* fluorescence spectroscopy and imaging for oncological applications. *Photo-chemistry and Photobiology*, 68(5):603–32, 1998.
27. J. Weston and C. Watkins. Multi-class support vector machines. *Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science*, 1998.