Available online at www.sciencedirect.com
ScienceDirect
Journal of Photochemistry and Photobiology B: Biology 85 (2006) 109–117
Journal of Photochemistry and Photobiology B:Biology
www.elsevier.com/locate/jphotobiol

Multi-class classification algorithm for optical diagnosis of oral cancer

S.K. Majumder [a,*], A. Gupta [b], S. Gupta [b], N. Ghosh [a], P.K. Gupta [a]

[a] *Biomedical Applications Section, Centre for Advanced Technology, Indore 452013, India*
[b] *Department of Computer Engineering, Netaji Subhash Institute of Technology, New Delhi 110075, India*

Received 24 April 2006; received in revised form 23 May 2006; accepted 31 May 2006
Available online 12 July 2006

Abstract

We report development of a direct multi-class spectroscopic diagnostic algorithm for discrimination of high-grade cancerous tissue sites from low-grade as well as precancerous and normal squamous tissue sites of human oral cavity. The algorithm was developed making use of the recently formulated theory of total principal component regression (TPCR). The in vivo autofluorescence spectral data acquired from patients screened for neoplasm of oral cavity at the Government Cancer Hospital, Indore, was used to train and validate the algorithm. The diagnostic algorithm based on TPCR was found to provide satisfactory performance in classifying the tissue sites in four different classes – high-grade squamous cell carcinoma, low-grade squamous cell carcinoma, leukoplakia, and normal squamous tissue. The classification accuracy for these four classes was observed to be ~94%, 100%, 100% and 91% for the training data set (based on leave-one-out cross-validation), and was ~90%, 90%, 85% and 88%, respectively for the corresponding classes for the independent validation data set.

Keywords: Autofluorescence spectroscopy; Multi-class classification; Diagnostic algorithm; Leukoplakia; Oral cancer; Total principal component regression (TPCR); Squamous cell carcinoma (SCC)

1. Introduction

A large body of work carried out over the last decade has established the applicability of optical spectroscopic technology for non-invasive, in situ, near-real time diagnosis of cancer [1–4]. Successful clinical application of this approach requires a suitable diagnostic algorithm that can rapidly and accurately classify the measured spectra from an unknown tissue making use of the stored database of spectra of tissues of known histopathologic classification. Over the years a variety of diagnostic algorithms of varying rigor have been developed for optical diagnosis of cancer [5–32]. One approach followed was to select empirically the discrimination indices from the observed differences in the spectral features. The empirically selected indices can be absolute or normalized fluorescence intensities [5–11], ratio of intensities at selected pairs of emission wavelengths [12–15] or ratio of integrated intensities over appropriately chosen wavelength bands [16]. These have been used directly for discrimination [5–8,12–15], or as inputs to statistical analytical techniques like multivariate linear regression (MVLR) analysis to form a discrimination function [9–11,16]. The more recent efforts are directed towards using statistical pattern recognition techniques to exploit the information content of the entire spectral data for extracting the best diagnostic features and use of these for accurately classifying tissue into corresponding histopathologic categories. Although linear techniques like Fisher's linear discriminant (FLD), principal component analysis (PCA), etc. [17–22] have been used successfully for algorithm development, the use of artificial neural network (ANN) [23–25], wavelet transforms [26], maximum representation and discrimination feature (MRDF) [27], and more recently support vector machine (SVM) [28–30] has been found to provide superior performance. Amongst

* Corresponding author. Present address: Department of Biomedical Engineering, Vanderbilt University, 5818 Stevenson center, Nashville, TN 37235, United States. Tel. +1 615 322 6245; fax: +1 615 343 7919.
E-mail address: shovan.k.majumder@vanderbilt.edu (S.K. Majumder).

1011-1344/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jphotobiol.2006.05.004

all these, SVM, in particular, is the best suited for this kind of supervised classification problems. The use of the theory of relevance vector machine (RVM), for development of a probability based algorithm for optical diagnosis of cancer has also been reported [31,32]. The information on the posterior probabilities of class membership is particularly important in clinical settings where the misclassification cost associated with some classes (false negative for cancer) may be significantly higher for that of others (false positive for cancer) [31,32].

Although the algorithms based on the above mentioned approaches have been found to provide excellent diagnostic performances, their major limitation is that most of them are inherently dichotomous, i.e., can be used only for binary classification, like cancerous vs. non-cancerous [5–22,26–32]. However, in clinical situation, in order to decide an appropriate mode of therapy, one often needs to classify the tissue site being interrogated into more than two classes. These classes may correspond to different pathological conditions or say grades of cancer. One approach to achieve this objective is to use multiple dichotomous classifiers in succession and heuristically combine them to solve such multi-class classification problem [33]. This approach, however, is not only non-optimized but also time-consuming. An alternate approach that could be considerably faster and therefore useful for clinical situation is to develop an algorithm that is inherently polychotomous, i.e., capable of simultaneously classifying spectral data into more than two classes. Although non-linear algorithms based on ANN [23–25] have the capability of direct multi-class classification and have also been used for that purpose by few researchers [23,24], the major disadvantage of ANNs is that they typically need a priori selection of a number of ad hoc parameters associated with the learning or the optimization technique to be used. Moreover, the weights in an ANN are trained iteratively which may lead to problems with convergence [34]. A novel linear technique that overcomes the limitations of ANN and doses not require a priori selection of ad hoc parameters is the technique of Total Principal Component Regression (TPCR), developed recently by Tan et al. [35] to classify various cancers based on gene expression profiles. The major advantage of TPCR is that unlike the iterative nature of non-linear ANN based algorithms, it provides a closed form expression of the discrimination function [35].

In this paper, we report development of a polychotomous spectroscopic diagnostic algorithm based on the theory of TPCR [35].

2. Materials and methods

2.1. Instrumentation

In vivo autofluorescence spectra were recorded using a N_2 laser (337 nm) based portable fluorimeter described earlier [22,32]. It comprised of a sealed-off pulsed N_2 laser, a spectrograph (Acton Research Corporation, 15 Discovery Way, Acton, MA 01720-4482, USA), an optical fiber probe and a gateable intensified CCD detector (4 Quik 05A, Stanford computer optics, Inc., 780 Cragmont Avenue, Berkeley, CA 94708, USA). The diagnostic probe, developed in-house, was a fiber bundle, which had two legs; one contained a single quartz fiber (400 μm core diameter, 0.22 NA) and the other contained six quartz fibers (400 μm core diameter, 0.22 NA). The two legs merged to form a common fiber bundle that consisted of a central fiber, surrounded by a circular array of six fibers. The central fiber delivered excitation light to the tissue surface and the six fibers surrounding the central fiber collected tissue fluorescence from the surface area directly illuminated by the excitation light. The proximal ends of the collection fibers were arranged in a vertical array and the light coming from the distal end was imaged at the entrance slit of the spectrograph coupled to the intensified CCD detector. The common end of the fiber bundle was enclosed in a stainless steel (SS) tube (9 mm outer diameter and 60 mm long). The tip of the probe was shielded by a quartz optical flat 2 mm thick to provide a fixed distance between tissues and the fibers for improved collection of fluorescence and also to protect contamination of the fiber tips with tissue fluids. The spectral data acquisition was computer controlled. The gateable ICCD camera was triggered using a trigger signal generated by sensing the current through the N_2 laser spark gap by a current transformer. The intrinsic delay between the trigger signal and the opening of the shutter was ~25 ns. In order to synchronize the ICCD shutter opening with the N_2 laser pulse induced tissue fluorescence the irradiation of tissue surface with the N_2 laser pulse was optically delayed by guiding the N_2 laser pulse through an optical fiber of appropriate length. The gate-width used was 100 ns. Since the decay times of fluorescence from different constituents of the tissue are typically <10 ns, it ensured that practically all the tissue fluorescence was collected.

2.2. Patient selection

The study involved 30 normal volunteers with no history of the disease of the oral cavity and 68 patients enrolled for medical examination of the oral cavity at the outpatient department (OPD) of the Government Cancer Hospital, Indore. Informed consent was obtained from each patient as well as the normal volunteers who participated in this study. Age, sex, and details of smoking habit (if any) were also recorded for all subjects included in the study. The age of the patients with cancer or leukoplakia ranged from 37 to 75 with mean age and standard deviation of 50 and 10, respectively, whereas for normal volunteers the age ranged from 24 to 58 with mean and standard deviation being 38 and 17. The overall ratio of male to female population was ~2 with the ratio for patients and for normal volunteers being ~1.5 and ~3.5, respectively. As far as tobacco habits are concerned, ~95% of the patients and ~60% of the normal volunteers had habits of either smoking or