The Bag of Words Approach for the Classification of Head and Neck Cancers Using Diffuse Reflectance Spectroscopy

Ahmed Karam Eldaly[1], Yannick Benezeth[1], Virginie Flaus[2], Christian Duvillard[2], Alexis Bozorg Grayeli[1,2], and Franck Marzani[1]

[1] LE2I UMR6306, CNRS, Arts et Métiers, Université de Bourgogne, Franche-Comté, F-21000 Dijon-France.
[2] Hopital de Dijon, Service ORL, F-21000 Dijon-France.
{ahmed.karam@h-eng.helwan.edu.eg}

Abstract. Diffuse Reflectance Spectroscopy (DRS) is a leading technique for the detection of head and neck cancers. It can capture information regarding tissue absorption and scattering. In this research work, we propose a novel method for the identification of normal and Squamous Cell Carcinoma (SCC) mucosa tissues using the Bag Of Words (BOW) approach. The study included 70 spectra from normal mucosa tissue sites and 70 spectra from SCC mucosa tissue sites. First, the spectra are preprocessed by extracting the useful wavelength range, denoising and reducing the inter and intra patient variability. Subsequently, features are extracted from each spectrum by continuously sliding a window with a pre-defined length along each spectrum to extract a group of local segments. Discrete Wavelet transform (DWT) is then employed for each segment. Next, we construct the codebook to represent each spectrum by a histogram of codewords at which each bin in the histogram is a count of a codeword appeared in the spectrum. Finally, the histogram representation is used as input for classification. The maximum accuracy reported is 94.28% with sensitivity and specificity of 91.42% and 97.14% respectively.

Keywords: Head and neck squamous cell carcinoma, Diffuse reflectance spectroscopy, Bag of words, Clustering, Codebook, Classification.

1 Introduction

The annual incidence of head and neck cancers worldwide is between 400,000 and 600,000 cases with around 223,000 and 300,000 deaths each year [9]. Cancer progression of the oral cavity starts by hyperplasia which is the benign stage of the cancer, dysplasia and finally Squamous Cell Carcinoma (SCC). About 90% of all Head and Neck cancers are Squamous Cell Carcinomas (HNSCC) which is the most malignant stage of the cancer. Most HNSCCs arise in the epithelial lining of the oral cavity, oropharynx, larynx and hypopharynx[10]. Alcohol and tobacco are known risk factors for most head and neck cancers, and incidence rates are found to be higher in regions with high rates of alcohol and tobacco consumption [6].

Currently, the gold standard technique for the detection and diagnosis of head and neck cancers is histopathology assessment of a biopsy of the tissue [3], however it is subjective to the views of the clinicians. Therefore, in order to ensure an accurate pathological diagnosis, a suspicious oral lesion may need multiple biopsies to avoid misdiagnosis of the most severe location. However, only a limited number of biopsies can be taken because of the invasiveness of the procedure.

Emerging non-invasive or minimally invasive techniques based on optical spectroscopy are showing great promise and will enable opportune diagnosis to improve patient cure and survival rates. Optical spectroscopy is the studying of the properties of the tissue based on analyzing how it interacts with light. The main advantages of using optical spectroscopy for the analysis of head and neck cancers are that it does not require tissue removal, not greatly affected by artifacts or sampling errors, and can provide quantitative information regarding tissue morphology and biochemistry that is largely free of subjective interpretation.

Diffuse Reflectance Spectroscopy (DRS) is one of the simplest spectroscopic techniques for studying biological tissues. It can provide biochemical information from the absorption (e.g. due to oxygenated haemoglobin) and structural information from scattering (e.g. nuclear size, epithelial thickness). In this work, motivated by the success of the bag of words approach for image analysis[4] and biomedical time series analysis[18], we propose a novel method based on the bag of words approach to automatically identify normal and SCC mucosa tissues since there are few approaches designed in the literature for this purpose and no approaches designed using this method.

The paper is organized as follows: section 2 reviews the literature for the identification of head and neck cancers, section 3 describes the proposed method, section 4 explains the experimental setup and data acquisition, section 5 gives a detailed analysis for the experimental results and finally, section 6 gives a conclusion about our work and proposes new research directions to continue this work.

2 Related Work

In literature, a wide variety of Computer Aided Diagnosis (CAD) systems to discriminate the different grades of head and neck malignancies using DRS have been reported. These CAD systems usually involve three main steps: pre-processing, feature extraction and selection and finally classification and evaluation.

1. **Pre-processing:** The pre-processing usually starts with the extraction of the useful wavelength range which most probably lies within the visible wavelength range i.e. 400 : 700 nm. After that, the spectra are passed to data denoising for noise suppression, smoothing and interpolation of missing data

and finally, normalization for the reduction of the inter and intra patient variability.

2. **Feature Extraction:** Four main categories of features have been observed in the literature regarding the classification of head and neck cancers: (1) dimensional reduction using Principle Component Analysis (PCA) which is the most widely used method [17][1] (2) spectral features which encode the change in the biological properties of the diseases since they give information about absorption and scattering at certain wavelengths [11], (3) model-based features which are concerned with the calculation of the absorption μ_s and scattering μ_a coefficients of the tissues by building models of these tissues and inversely calculating these coefficients [15] and finally (4) hybrid features which is either a mixture of spectral and model-based features [14] or using different spectroscopic techniques [12].
3. **Classification:** The classification step is used to classify the spectra into either normal or abnormal. The most commonly used classifier for the diagnosis of head and neck cancers using DRS is Linear Discriminant Analysis (LDA) [16][8].

3 Methodology

Basically, the proposed method is based on the bag of words approach to automatically identify normal and SCC mucosa tissues. First, we continuously slide a window with a pre-defined length along each spectrum to extract a group of local segments. Discrete Wavelet Transform (DWT) is then employed for each segment. Next, similar to the bag of visual words model of image analysis, all of the local features from the training dataset are clustered using the K-means algorithm to create a codebook. The cluster centres are treated as codewords. Then, a local segment is assigned the codeword that has the minimum distance to it. Each spectrum is then represented as a histogram of codewords at which each bin is a count of a codeword appeared in the spectrum. Finally, the BOWs representation is used as input for classification. Figure 1 shows a schematic representation for the proposed methodology.

3.1 Pre-processing

The pre-processing of the spectra includes three main steps:

- **Extraction of the Useful Wavelength Range:** Since the spectra are measured by a wide range source (347.7 nm : 1100.5 nm), checking the spectra visually shows that the useful wavelength range lies between 445 nm and 960 nm. The spectra are very noisy below and above this wavelength range. The total number of points of each spectrum after extracting the useful wavelength range is 898 points.

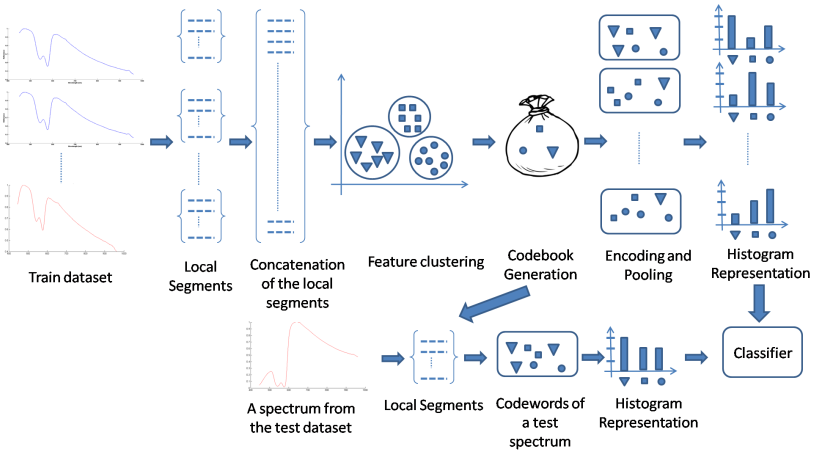

Fig. 1: A schematic representation of the Bag Of Words (BOW) approach.

- **Averaging:** An averaging window of size $1 * 5$ *points* which corresponds to $1 * 3$ *nm* length is employed for each spectrum.
- **Normalization:** In this work, each spectrum is normalized by its peak intensity. This ensures that the range of reflectances for all of the spectra lies between 0 and 1.

Figure 2 shows the result of applying the pre-processing steps described above on four spectra included in the study.

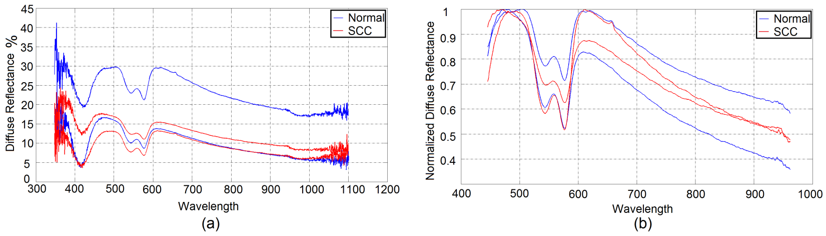

Fig. 2: The pre-processing of the spectra: (a) Original spectra, (b) Final result after pre-processing.

3.2 Feature Extraction

Discrete Wavelet Transform (DWT) is a common feature extraction method for the classification of skin cancers using diffuse reflectance spectroscopy [5]. In this work, we test DWT as a feature extraction method with the bag of words approach. There are many types of mother wavelets which can be used for wavelet analysis. Different mother wavelets used to analyze the same signal will produce different results [13]. Ngui et al. [13] suggested to select the mother wavelet based on the similarity with the signal being processed. Following this approach, different wavelets were tested which are near in the similarity to the wavelength band 520 : 600 nm such as db4, db5, db6, sym4, sym5 and sym6 where ’db’ and ’sym’ refer to ’Daubechies’ and ’Symlet’ respectively.

3.3 Local Segment Extraction

After the pre-processing of the spectra, we continuously slide a window with a pre-defined length along each spectrum and extracting a group of local segments. Each local segment is then transformed into wavelet domain. In this work, a single level DWT is employed to decompose a local segment into approximate and detail coefficients. The approximate coefficients are then used as a feature vector to represent each local segment since the detail coefficients are almost zeros which means that there is no high frequency changes. All of the local segments extracted from all of the spectra in the train dataset are concatenated for the generation of the codebook.

3.4 Codebook Generation

The codebook is a set of words which are also called codewords. The codebook is generally created by performing clustering of the training data. The K-means clustering algorithm [4] is commonly used to construct the codebook. Similar to the codebook construction in image and video analysis, we cluster all the local segments from the spectra of the training dataset to construct the codebook. The clustering centres estimated by the K-means clustering are regarded as basis elements of the codebook, i.e., codewords.

3.5 Feature Encoding and Pooling

The purpose of the encoding step is to assign each feature vector of each local segment to the nearest word in the codebook. As a result a single vector P containing the corresponding words of each feature descriptor is obtained. The pooling step is simply performed by counting the number of occurrences of each word in the resultant vector P, and then normalizing this vector using L2-Normalization.

3.6 Classification

Three classifiers are tested in this work: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) and K-Nearest Neighbour (KNN). The Matlab built-in functions for KNN and LDA are used and the LIBSVM [2] library is used for SVM. The classification results are obtained by carrying out 10 fold cross validation at which each fold includes 7 normal and 7 SCC spectra. The system performance is evaluated by building the confusion matrix for each tested fold and summing all the confusion matrices for all of the tested folds and then calculating the sensitivity, specificity and accuracy.

4 Experimental Setup and Data Acquisition

A portable spectrometer was used for the measurements of diffuse reflectance spectra from head and neck mucosa. Figure 3 shows a block diagram for the different parts of the device. The system consists of a compact tungsten halogen lamp (Avantes Optics, model: AvaLight-Hal) for tissue illumination and a fiber optic spectrometer (Avantes Optics, model: AvaSpec-2048L VIS-NIR) connected to a USB port of a laptop for recording the diffuse reflectance spectra from the tissues. The central fiber is bifurcated into six fibres maintained in one tube for tissue illumination and one fiber to collect the diffusely reflected light emissions. The reflection probe tip is terminated in a stainless steel ferrule of 10 cm long and 1.5 mm in diameter for easy access to interior areas and to facilitate sterilization before and after use. The AvaSoft 8 software was then used to record the diffuse reflectance measurements in the range of 347.7 : 1100.5 nm. The resolution of measurements is 0.598 nm which means that each spectrum is composed of 1315 reflectance points. To maintain hygiene, a disposal plastic sleeve was inserted at the probe tip for each subject included in the study, furthermore the spectrometer was also being sterilized before each acquisition.

The study included 8 healthy volunteers from which 40 spectra were acquired from normal mucosa tissue sites and 20 patients from which 30 spectra were acquired from normal mucosa tissue sites and 70 spectra from SCC mucosa tissue sites. Biopsy specimens were taken from the SCC mucosa tissue sites to get the ground truth.

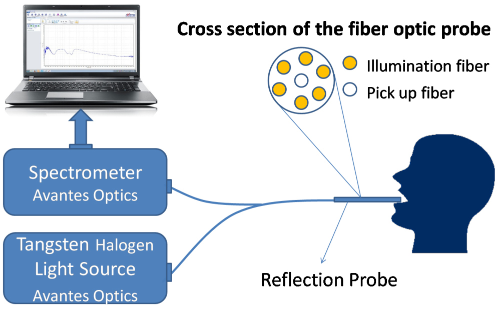

Fig. 3: Schematic of the diffuse reflectance point monitoring set-up used in the clinical trial.

5 Results and Discussion

Figure 4 (a) shows a plot of the mean spectral intensity of the 40 spectra from the 8 healthy volunteers and the 30 spectra from normal mucosa tissue sites from the 20 patients and the corresponding standard deviation at each point. We can observe that there is a large variation for the normal data from the patients and the variation is smaller for the normal data from the healthy volunteers in the wavelength band of 445 : 590 nm. The variation then becomes nearly equal in the wavelength band of 590 : 960 nm. We assume that the variation between the two types of data refers to the field cancerization effect [7] in which the normal tissues are affected by the cancerous cells. In this study, we assumed that the proposed approach should identify normal mucosa for both healthy people and patients, so we used normal data from both. Figure 4 (b) shows the mean and the standard deviation of the SCC spectra and all of the normal spectra from both healthy volunteers and patients. We can observe that there is a large variation between the two types of data over the entire wavelength range.

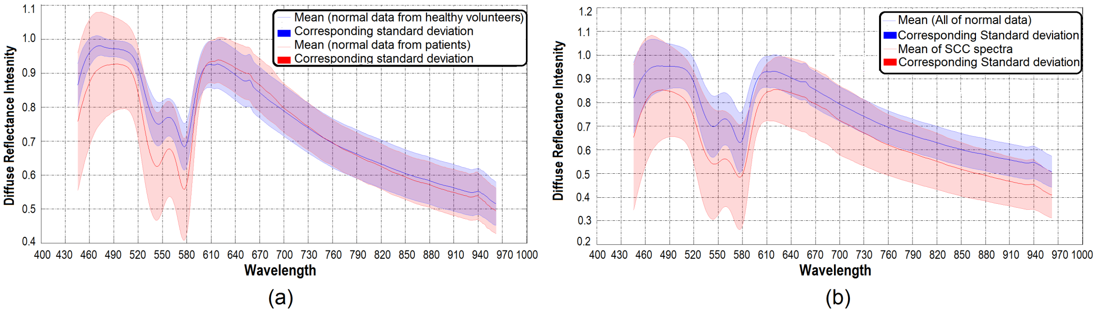

Fig. 4: Mean spectral intensity (solid line) and the corresponding standard deviation (shaded area) at each point for (a) 40 spectra from normal mucosa tissue sites from healthy volunteers and 30 normal mucosa tissue sites from patients. (b) All normal mucosa spectra from both healthy volunteers and patients (70 spectra) and 70 spectra from SCC mucosa tissue sites.

5.1 Classification based on dimensional reduction using PCA

Classification based on dimensional reduction of the spectra using PCA is the mostly widely used method in the literature as we discussed in section 2 . In this work, this method was implemented for comparison purposes. The maximum accuracy reported is 90.71% at number of principle components of 21 and LDA classification, the corresponding sensitivity and specificity are 82.57% and 98.57% respectively. We can observe that there is a high misclassification rate for SCC mucosa tissues.

5.2 Classification based on the BOW approach

Since each spectrum is divided into a group of local segments, three segment lengths were tested, 24 nm, 48 nm and 78 nm, the corresponding number of extracted features for each spectrum is 22, 11 and 6 respectively.

Classifiers Performance:
Figure 5 shows a plot of the accuracy versus different numbers of clusters for db6 wavelet for segment lengths of 24 nm, 48 nm and 78 nm. We observed that for all of the wavelets and all of the segment lengths, the KNN classifier has the highest performance. The LDA and SVM classifiers have near performance, however the

LDA performance decreases with high numbers of clusters for segment lengths of 24 nm and 48 nm.

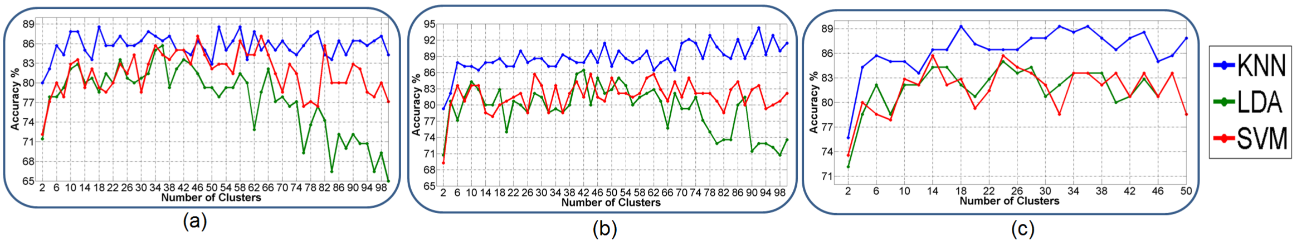

Fig. 5: A plot of the number of clusters versus the accuracy for db6 wavelet for segment lengths of (a) 24 nm, (b) 48 nm and (c) 78 nm.

Effect of Segment Lengths and Wavelets:
Figure 6 shows a plot of the three tested segment lengths versus the maximum accuracy achieved by each wavelet for the three tested classifiers. The number of clusters corresponding to each maximum accuracy is shown above each bar. For KNN, we can observe that the segment length 48 nm has the best performance since it gives the highest accuracies for all of the wavelets compared to the other two segment lengths. The maximum accuracy obtained is 94.28% for the db5 and db6 wavelets. The corresponding sensitivities and specificities are 90% and 98.57% respectively for db5 and 91.42% and 97.14% respectively for db6. However, we can assume that the best performance is for db6 since it gives higher sensitivity than db5. If a patient with SCC mucosa tissue is misdiagnosed as normal (false negative), that means we are in troubles. However, if a patient with normal mucosa tissue is misdiagnosed as SCC mucosa tissue (false positive), he or she may repeat the test to make sure about the disease. False negative is more pernicious than false positive. The maximum accuracy obtained for segment length of 24 nm is 89.28% for db4, db5, sym4 and sym6. Segment length of 78 nm has a maximum accuracy of 92.14% corresponding to db5 wavelet. The segment length of 78 nm has better performance than 24 nm for all of the wavelets except for sym4 which gives equal accuracies.

We can observe that there is no specific rule for the performance of the LDA and SVM classifiers since all of the wavelets are behaving differently for each segment length.

6 Conclusion and Future Work

In this work, diffuse reflectance spectroscopy was used to acquire spectra from normal and squamous cell carcinoma mucosa tissues from the head and neck. The spectra were represented as histograms of codewords using the bag of words approach. The features were extracted by dividing each spectrum into segments of a pre-defined length. Discrete wavelet transform was then used to transform the segments to the wavelet domain. We could achieve an accuracy of 94.28%,

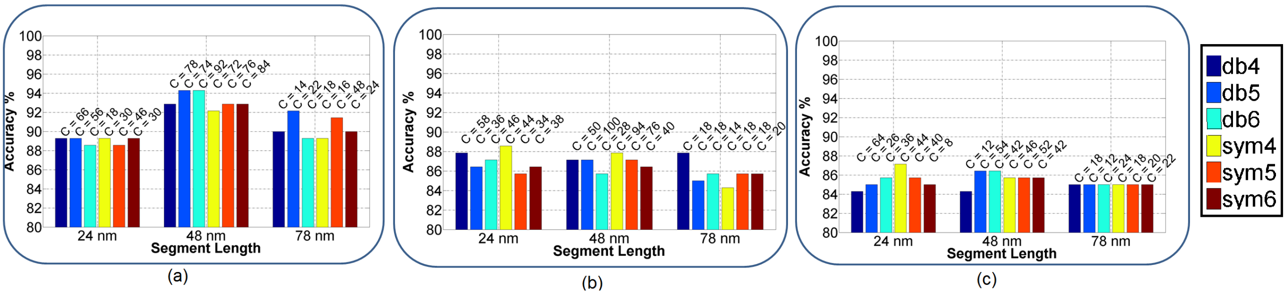

Fig. 6: Maximum accuracy obtained by each wavelet for each segment length using (a) KNN, (b) SVM and (c) LDA classification. The corresponding number of clusters is shown above each bar.

the corresponding sensitivity and specificity are 91.43% and 97.14% respectively. In comparison with the literature, classification based on dimensional reduction of the spectra using PCA was employed. The maximum accuracy achieved was 90.71%, the corresponding sensitivity and specificity are 82.28% and 98.57% respectively. We can observe that the classification based on the bag of words approach could achieve higher sensitivity than the classification based on dimensional reduction of the spectra using PCA with keeping near specificity. Furthermore, Muller et al. [12] used three spectroscopic techniques (intrinsic fluorescence, diffuse reflectance and light scattering) for the discrimination between normal and squamous cell carcinoma lesions in the oral cavity. The diagnosis of each site was determined by assigning the result agreed by at least two spectroscopic techniques from the three. the maximum sensitivity and specificity achieved were 96% and 96% respectively. To conclude, diffuse reflectance spectroscopy provides useful diagnostic information about the morphological and biochemical changes in head and neck normal and cancerous mucosa tissues. Also, the bag of words approach with discrete wavelet transform gives promising results for the discrimination between normal and squamous cell carcinoma mucosa tissues compared with the literature. As a future work, we propose to acquire more data and performing more experiments, testing the performance of combining diffuse reflectance spectroscopy with other spectroscopic techniques, namely autofluorescence and light scattering spectroscopy and finally, extending the proposed approach to deal with more challenging cancer grades, namely hyperplasia and dysplasia.

References

1. Beattie, J.R., Glenn, J.V., Boulton, M.E., Stitt, A.W., McGarvey, J.J.: Effect of signal intensity normalization on the multivariate analysis of spectral data in complex real-worlddatasets. Journal of Raman Spectroscopy 40(4), 429–435 (2009)
2. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

3. Clara, S.: Histological typing of cancer and precancer of the oral mucosa
4. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 2, pp. 524–531. IEEE (2005)
5. Garcia-Uribe, A., Zou, J., Duvic, M., Cho-Vega, J.H., Prieto, V.G., Wang, L.V.: In vivo diagnosis of melanoma and nonmelanoma skin cancer using oblique incidence diffuse reflectance spectrometry. Cancer research 72(11), 2738–2745 (2012)
6. Hashibe, M., Brennan, P., Chuang, S.c., Boccia, S., Castellsague, X., Chen, C., Curado, M.P., Dal Maso, L., Daudt, A.W., Fabianova, E., et al.: Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the international head and neck cancer epidemiology consortium. Cancer Epidemiology Biomarkers & Prevention 18(2), 541–550 (2009)
7. Jayanthi, J., Mallia, R.J., Shiny, S.T., Baiju, K.V., Mathews, A., Kumar, R., Sebastian, P., Madhavan, J., Aparna, G., Subhash, N.: Discriminant analysis of autofluorescence spectra for classification of oral lesions in vivo. Lasers in surgery and medicine 41(5), 345–352 (2009)
8. Jayanthi, J., Nisha, G., Manju, S., Philip, E., Jeemon, P., Baiju, K., Beena, V., Subhash, N.: Diffuse reflectance spectroscopy: diagnostic accuracy of a non-invasive screening technique for early detection of malignant changes in the oral cavity. BMJ open 1(1), e000071 (2011)
9. Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D.: Global cancer statistics. CA: a cancer journal for clinicians 61(2), 69–90 (2011)
10. Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Thun, M.J.: Cancer statistics, 2007. CA: a cancer journal for clinicians 57(1), 43–66 (2007)
11. Mallia, R., Thomas, S.S., Mathews, A., Kumar, R., Sebastian, P., Madhavan, J., Subhash, N.: Oxygenated hemoglobin diffuse reflectance ratio for in vivo detection of oral pre-cancer. Journal of biomedical optics 13(4), 041306–041306 (2008)
12. Müller, M.G., Valdez, T.A., Georgakoudi, I., Backman, V., Fuentes, C., Kabani, S., Laver, N., Wang, Z., Boone, C.W., Dasari, R.R., et al.: Spectroscopic detection and evaluation of morphologic and biochemical changes in early human oral carcinoma. Cancer 97(7), 1681–1692 (2003)
13. Ngui, W.K., Leong, M.S., Hee, L.M., Abdelrhman, A.M.: Wavelet analysis: mother wavelet selection methods. Applied mechanics and materials 393, 953–958 (2013)
14. Nieman, L.T., Kan, C.W., Gillenwater, A., Markey, M.K., Sokolov, K.: Probing local tissue changes in the oral cavity for early detection of cancer using oblique polarized reflectance spectroscopy: a pilot clinical trial. Journal of biomedical optics 13(2), 024011–024011 (2008)
15. Palmer, G.M., Ramanujam, N.: Monte carlo-based inverse model for calculating tissue optical properties. part i: Theory and validation on synthetic phantoms. Applied optics 45(5), 1062–1071 (2006)
16. Péry, E., Blondel, W.C., Tindel, S., Ghribi, M., Leroux, A., Guillemin, F.: Spectral features selection and classification for bimodal optical spectroscopy applied to bladder cancer in vivo diagnosis. Biomedical Engineering, IEEE Transactions on 61(1), 207–216 (2014)
17. de Veld, D.C., Skurichina, M., Witjes, M.J., Duin, R.P., Sterenborg, H.J., Roodenburg, J.L.: Clinical study for classification of benign, dysplastic, and malignant oral lesions using autofluorescence spectroscopy. Journal of biomedical optics 9(5), 940–950 (2004)
18. Wang, J., Liu, P., She, M.F., Nahavandi, S., Kouzani, A.: Bag-of-words representation for biomedical time series classification. Biomedical Signal Processing and Control 8(6), 634–644 (2013)