Smoking status classification by optical spectroscopy and partial least square regression

Audrey Kah Ching Huong¹*, Wan Mahani Hafizah Wan Mahmud¹, Kim Gaik Tay¹, Xavier Toh Ik Ngu¹²

¹Faculty of Electric and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia
²Center for Applied Electromagnetics, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia

*audrey@uthm.edu.my

Abstract. Smoking status of individuals is often revealed through self-reported data and interviews. The incidence of false reports severely impairs the proper assessment of the individuals’ health conditions and their risk to tobacco associated diseases, delays clinical intervention and treatment services. This paper presents the use of optical technique combined with partial least square (PLS) regression model in the classification of smoking status. The focus of this work is on light absorbance signals (by skin tissue) in the wavelength range of 520 – 600 nm; based on these data dendogram and PLS techniques are used to estimate the number of likely clusters within the considered dataset and to predict smoking status of individuals, respectively. The results from the processing of spectral information of smoking and nonsmoking populations revealed a high misclassification rate of 26.67 % using dendrogram method, but a considerably high accuracy of 90 % evaluated via leave one out cross validation was obtained using PLS component number 4. This study concluded that the spectral oscillation patterns and descending rates corresponded to nonsmoking and smoking individuals could be differentiated and specified using PLS technique in the determination of smoking status.

1. Introduction

Identifying an individual’s smoking status has become an increasingly important part of public health’s responsibility in this era of civilization, catalysed by the increased environmental awareness and high level of health education among the general public. It was reported by World Health Organization in year 2018 that 14.6 % of population worldwide are smokers [1], these statistics were shown to be affected by a variety of factors such as demographic characteristics, socioeconomic and education level of an individual [2], the revolution in broadcasting industries such as advertisement, film, television and media, and local tobacco control policies [3]. Tobacco use is known as a risk factor for numerous chronic diseases such as cardiovascular and respiratory diseases [4], gene defects, cancers and stroke. Other effects of tobacco use include premature birth and mortality.

For this reason, world leaders, independent organizations and economic major players from different countries are working together towards tobacco-free world by year 2040 [5]. This is evident with the cessation of tobacco production by Dunhill, a widely recognizable name in tobacco industry, and the passing of anti-smoking policy on the discontinuance of new smoking recruits in a Japanese university [6], all in the efforts to mitigate smoking habit and tobacco addiction among smokers. In addition, bans on tobacco advertising, promotion and sponsorship, taxes on tobacco products and...
graphic pictorial warning to educate and encourage awareness among the smokers are some examples of anti-smoking strategies undertaken in the global level.

The common proxy measures of identifying smokers including through diagnosis of tobacco-use disorder, counselling visits records, anti-smoking prescription history [7], and records of mobile apps that support smoking cessation [8]. However, these data may not be available for those who have not taken initiative to receive intervention for their smoking habit. Meanwhile the validity of records based on self-report approach is subjected to authentication as workers in [9] found a high percentage (of 48 %) of their recruits made false statements about their smoking status.

A rapid and accurate means of determining smoking status of an individual is of the utmost importance to identify the health condition of the person for applications such as in epidemiological studies of smoking, clinical studies of smoking related diseases [4], for effective treatment strategy, and for the evaluation of health insurance coverage. Among the measurable parameters related to tobacco use are traces of nicotine and increase of carbon monoxide (CO) level in blood. The state of the art of noninvasive technologies for the prediction of smoking status include pulmonary function test, urinary cotinine concentration test and exhaled carbon monoxide (eCO) tester, but there are reservations about the sensitivity and specificity of these technologies. Previous reports [9] found contradictory readings obtained using these technologies in a clinical trial involving 144 subjects.

Optical system that uses light waves of different wavelengths and a light sensor, or its advanced forms (e.g. polarization technique) [10], to noninvasively extract the underlying health conditions has seen many of its clinical applications; these are such as in the diagnosis of hyperbilirubinemia [11], hyperoxia [12] and CO poisoning [12, 13]. The imaging capacity of this technology is inherently limited to the skin penetration depth of a few hundred microns (i.e. dermal layers of human skin). Existing commercial available optical sensors to determine CO poisoning include Pulse CO-Oximetry (Masimo, USA), which operation is based on changes of the absorption spectra of hemoglobin derivatives.

Meanwhile with the recent massive growth and progression in the processing and memory capability of a computing unit, artificial intelligence (AI) plays a crucial role in nearly all systems or technologies that required decision making. It is, therefore, the aim of this study to explore the use of optical technology combined with AI system as a means for rapid classification of a person’s smoking status based on their blood CO level.

2. Materials and methods

2.1. Spectral data sample

Thirty light absorbance spectra recorded from right index finger of 18 smoking and 12 nonsmoking subjects in [12] were considered in this study. In this preceding research, light reflected from the examined site was spectrally resolved by a spectrometer (model no. USB4000, OceanOptics, Inc) across the wavelength range of 200 – 1100 nm at a resolution of 0.2 nm; the detected light signals was used in the analysis to predict blood oxyhemoglobin and carboxyhemoglobin saturations. The smoking characteristics of 18 smokers are listed in Table 1. The subject index of nonsmoking control group is from 19 to 30 (aged between 20 and 38 years). The difference in the age range of subjects in these two groups was due to the limited availability in non-smoking subjects of other age categories who were willing to participate in the study.

This research considered only data across wavelengths from 520 to 600 nm based on the previous reports of the distinctive optical signature of hemoglobin components, specifically oxyhemoglobin, deoxyhemoglobin and carboxyhemoglobin, across this wavelength range (see reports of [14]). The use of tobacco would increase blood CO level (i.e. carboxyhemoglobin level) rendering changes in light absorbance spectra. This study used unsupervised dendrogram method to provide an estimate of grouping of dataset and a visual summary of similarity among spectra. The latter was inferred based on the height of the link in the dendrogram. The supervised regression analysis was via PLS component number from 2 to 4; the training and testing data were iteratively chosen based on the concept of leave one out cross validation.
Table 1. The information on age and smoking characteristics of smoking individuals considered in this study (Subject index 1–18).

| Subject index | Age | Daily tobacco use (Cigarette sticks) | Number of years smoking |
|---------------|-----|-------------------------------------|-------------------------|
| 1             | 22  | 9                                   | 5                       |
| 2             | 28  | 8                                   | 10                      |
| 3             | 20  | 5                                   | 2                       |
| 4             | 21  | 5                                   | 3                       |
| 5             | 52  | 12                                  | 37                      |
| 6             | 38  | 10                                  | 20                      |
| 7             | 20  | 14                                  | 8                       |
| 8             | 43  | 12                                  | 20                      |
| 9             | 25  | 2                                   | 2                       |
| 10            | 20  | 7                                   | 7                       |
| 11            | 23  | 10                                  | 5                       |
| 12            | 48  | 20                                  | 20                      |
| 13            | 43  | 20                                  | 25                      |
| 14            | 58  | 20                                  | 15                      |
| 15            | 62  | 36                                  | 50                      |
| 16            | 51  | 4                                   | 20                      |
| 17            | 24  | 20                                  | 10                      |
| 18            | 24  | 5                                   | 3                       |

2.2. Partial Least Square Regression Discriminant Analysis (PLS-DA)

This study explored the use of supervised PLS-DA technique in the identification of smoking status of a person using MATLAB (version 2018a). This method predicts the responses, $Y$, in a population via extracting latent variables from sampled factors and responses as followed [15]:

\[
X = T * P' + E_0 \\
Y = U * Q' + F_0
\]  \hspace{1cm} (1)

where $X$ and $Y$ are the predictor and response, respectively, $T$ and $U$ are score matrices or latent variables, $P'$ and $Q'$ are loading matrices while $E_0$ and $F_0$ are residuals. The scores are solved via least square regression as followed:

\[
U = TB + F_1
\]  \hspace{1cm} (2)

where $B$ and $F_1$ are inner relation coefficient and residuals, respectively, to give the overall partial regression model in formula (3).

\[
\hat{Y} = X * (P * B * Q') + F.
\]  \hspace{1cm} (3)

$\hat{Y}$ is a categorical variable represented by 0 (to indicate nonsmoker) and 1 (for smoker). The response loadings with $\hat{Y} \geq 0.5$ were discriminated as that of a smoker.

3. Results

Light absorbance spectra for smoking and nonsmoking subjects are shown in Figure 1. The dendrogram in Figure 2 shows three distinct groups. While three (nonsmoking) subjects of index 20, 24 and 30 were misclassified, five spectral data, from that of both smoking and nonsmoking groups (index 4, 10, 16, 21, 25), were found in the third cluster. The bottom diagrams of Figure 2 are the spectral data of subject index 3, 23 and 21 from the clustered groups. The subjects’ smoking status
value predicted using different PLS component number is shown in Figure 3, also shown in this diagram is a dotted horizontal line drawn to indicate the set threshold value of 0.5. The accuracy of 76.67 %, 93.3 % and 90 % was achieved using PLS component number 2, 3 and 4, respectively.

![Figure 1](image1.png)

**Figure 1.** Light absorbance of smoking and non-smoking individuals.

![Figure 2](image2.png)

**Figure 2.** Cluster analysis of spectral data for 30 subjects. (Bottom left to right) Spectral data of subject index 3, 23 and 21 from three distinct groups.

4. Discussion
The dendrogram in Figure 2 showed that the absorbance spectrum for most of the subjects fall into the two distinct clusters, distinctive difference in the visibility of oscillation patterns can be observed in their spectral graphs in the wavelength range of 530-580 nm. Meanwhile several individuals (subject index 4, 10, 16, 21 and 25) from both control and investigation groups were categorized under the third class producing misclassification rate of 26.67 %. An investigation into the absorbance spectrum of these individuals revealed comparative higher magnitudes of light absorbance for individuals in this
group, suggesting skin pigmentation as an important feature in spectral classification. It is interesting to note on the striking similarity in the data of several smoking individuals (e.g. subject index 11 and 17). It was found that despite similar absorbance spectrum was measured, their smoking habit and history in Table 1 are relatively different. The main reason for this remained to be clarified, but the authors have not ruled out the possibility of prevalence monotonic effects of CO in blood for smoking individuals.

![Figure 3](image.png)

**Figure 3.** Smoking status value predicted for different subjects using partial least square (PLS) component number from 2 to 4. Value greater than 0.5 was classified as positive.

Even though this study has not considered the effects of variations in nicotine content of cigarette smoke, the lapse of time between the last cigarette smoked and the time of experiment, and the different ability of one’s body to expel the CO, the results in Figure 3 showed remarkable consistency in the value predicted for smoking subjects using PLS regression of 4th component. The DA results revealed a slightly lower accuracy of 90 % as compared to that using PLS component number 3, which produced an accuracy of 93 %. The two peaks in the light absorbance spectra, contributed by oxyhemoglobin component, gradually resolved in the investigation group in Figure 1 as compared to that of the control group. This work found that the absorbance spectrum of subject index 6 and 12, which were marginally misclassified in Figure 3, has similar oscillation patterns and magnitudes as that of other smokers, the most apparent disparity is in the spectrum descending rate in between wavelength 580 and 600 nm. The descending pattern observed for these two spectra shows moderate downward movement, it is neither extremely fast nor slow, and is of similar value to that of subject index 27 and 29. Interestingly this misclassification was not observed in subject index 18, whose spectral pattern is considerably similar to that of 12 in Figure 2, however this study has not eliminated the possibility of this being the source of the error.

This study is subjected to limitations including the small number of sample size and the gap in the information of the investigation group, but the results of this study could have implication for future healthcare management and intervention practices. Addressing the above mentioned limitations could be the next steps of this research, which comprises implementation of unsupervised clustering of spectroscopy data before using PLS-DA technique in the prediction of smoking status for new set of data.

5. Conclusion
This work demonstrated the use of PLS-DA in the prediction of a person’s smoking status, and a considerably good accuracy of 90 % was obtained using this technique. The findings of this work add to the literature that the prediction using this approach is independent of nicotine content of cigarette use and the last time a person smoke. This study found that, in addition to CO level, skin
pigmentation, spectral oscillation patterns and descending rates were among the important determinants in the prediction of smoking status.

Acknowledgments
This work is financially supported in part by the Ministry of Education Malaysia under FRGS (grant no. 1581), PRGS (grant no. K038) and Universiti Tun Hussein Onn Malaysia (B084, U848, H302, H310).

References
[1] World Health Organization WHO 2018 Tobacco Retrieved from https://www.who.int/news-room/fact-sheets/detail/tobacco
[2] Wang Y, Sung H Y, Yao T, Lightwood J and Max W 2017 Factors associated with short term transitions of nondaily smokers: socio-demographic characteristics and other tobacco product use J. Addict. 112 864-872
[3] Pomerleau J, Gilmore A, McKee M, Rose R and Haerpfer C W 2004 Determinants of smoking in eight countries of the former Soviet Union: results from the Living Conditions, Lifestyles and Health Study J. Addict. 99 1577–85
[4] Perez-Warnishner M T, Carballosa de Miguel M P and Seijo L M 2019 Tobacco Use Worldwide: Legislative Efforts to Curb Consumption Ann. Glob. Health 85 9
[5] Beaglehole R, Bonita R, Yach D, Mackay J and Reddy K S 2015 A tobacco-free world: a call to action to phase out the sale of tobacco products by 2040 Lancet 385 1011-1018
[6] Victor D 2019 Japanese university says it wont hired teachers who smoke Retrieved from The New York Times, https://www.nytimes.com/2019/04/25/world/asia/smoking-teachers-nagasaki-japan.html
[7] Desai R J, Solomon D H, Shadick N, Iannaccone C and Kim S C 2016 Identification of smoking using Medicare data- A validation study of claims based algorithm Pharmacoepidemiol Drug Saf. 25 1-8
[8] Heminger C L, Schindler-Ruwisch J M and Abroms L C 2016 Smoking cessation support for pregnant women: role of mobile technology Subst. Abuse Rehabil. 7 15-26.
[9] Stelmach R, Fernandes F L, Carvalho-Pinto R M, Athanozio R A, Rached S Z, Prado G F and Cukier A 2015 Comparison between objective measures of smoking and self-reported smoking status in patients with asthma or COPD: are our patients telling us the truth? J. Bras. Pneumol. 41 124-32.
[10] Huong A, Yip J and Ngu X 2017 Polarized optical system for noncontact imaging of epidermal microcirculation activities ACM Proc. Int. Conf. on Imaging, Signal Processing and Communication (Penang, Malaysia) p 59-62.
[11] Cheng N Y, Lin Y L, Fang M C, Lu W H, Yang C C and Tseng S H 2019 Noninvasive transcutanous bilirubin assessment of neonates with hyperbilirubinemia using a photon diffusion theory-based method Biomed. Opt. Express 10 2969-2984.
[12] Huong A, Philimon, S and Ngu X 2015 Non-invasive estimation of blood oxyhemoglobin and carboxyhemoglobin saturations using cumulant based forward model APRN J. of Eng. Appl. Sci. 10 8421.
[13] Jędrzejewska-Szczereska M, Karpienko K, Wróbel M S, Tuchin V V 2016 Biosensors for Security and Bioterrorism Applications. Advanced Sciences and Technologies for Security Applications ed D Nikolelis and G P Nikoleli (Switzerland: Springer)
[14] Zijlstra W G, Buursma A and van Assendelft O W 2000 Visible and near infrared absorption spectra of human and animal haemoglobin: determination and application (Netherlands:VSP)
[15] Chiang Y Q, Zhuang Y M and Yang J Y 1992 Optimal Fisher discriminant analysis using the rank decomposition Pattern Recognit. 25 101-111