Classification of human skin Raman spectra using multivariate curve resolution (MCR) and partial least squares discriminant analysis (PLS-DA)

I Matveeva¹, Y Khristoforova¹, A Moryatov², O Myakinin¹, I Bratchenko¹, S Kozlov² and V Zakharov¹

¹Samara National Research University, Moskovskoe Shosse 34, Samara, Russia, 443086
²Samara State Medical University, 89 Chapaevskaya street, Samara, Russia, 443099

E-mail: m-irene-a@yandex.ru

Abstract. The main purpose of the paper is classification of the human skin Raman spectra using partial least squares discriminant analysis (PLS-DA) into classes depending on the disease. In vivo Raman spectra of normal skin, basal cell carcinoma, malignant melanoma and pigmented nevus are considered. A feature of the approach is the analysis not of the Raman spectra themselves, but of the concentrations of the eight most significant spectra components identified using multivariate curve resolution (MCR). As a result, the ROC curve was calculated and the optimal classification threshold was chosen. The accuracy of the classification models ranged from 63.3 to 86.7%, depending on the model. The findings suggest that this approach could also be useful for classification of specific diseases.

1. Introduction
Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020. According to the World Health Organization, one of the most common in 2020 (in terms of new cases of cancer) was skin cancer [1]. While basal cell carcinoma (BCC) is the most common type of non-melanoma skin cancer and accounts for about 80% of the total number of cases, malignant melanoma (MM) is the most dangerous type of skin cancer. For example, in 2020, there were about 6,800 deaths from melanoma in the US [2]. It is well known that cancer mortality can be reduced by early detection of cases [1].

In recent years, optical methods such as Raman spectroscopy have been increasingly used to diagnose skin cancer [3]. Raman spectra of biological tissues are specific and can be used for successful differentiation of various pathologies in biological tissues. The possibility of in vivo tests and non-invasiveness make optical methods and, in particular, Raman spectroscopy, relevant for the analysis of skin diseases. During the development of a pathological neoplasm, a change in the biochemical composition of biological tissue occurs, which leads to a change in the Raman spectra [4]. Comparison of the Raman spectra of normal skin and different types of tumours shows the possibilities of their differential diagnosis [5].
However, the analysis of experimental Raman spectra is difficult due to the fact that the spectra contain a lot of information about all substances that make up the human skin [4, 5]. In this paper we propose a new approach to the analysis of experimental Raman spectra. Its essence is the preliminary unmixing the spectrum into components using multivariate curve resolution (MCR) and the subsequent analysis of the components by the partial least squares discriminant analysis (PLS-DA).

2. Materials and methods

2.1. Experimental data
We used in vivo Raman spectra of normal skin, basal cell carcinoma (BCC), malignant melanoma (MM) and pigmented nevus (PN). The Raman spectra was recorded using a portable spectroscopic setup which includes a thermally stabilized LML-785.0RB-04 diode laser module as an excitation source (785±0.1 nm central wavelength, 200 mW laser power) and a QE 65 Pro spectrometer (OceanOptics, Inc., USA) with CCD detector operating at -15 °C [6]. The registration of spectra using this setup was carried out in 800 – 1000 nm with 0.2 nm spectral resolution that corresponds to the 240 – 2236 cm⁻¹.

Then the spectra were cut in the range from 860 to 920 nm that corresponds to 1114 – 1874 cm⁻¹. The spectra cropped were preprocessed with baseline removal, and then smoothing by the Savitzky-Golay method, data normalization and centering, which were automatically applied in the “TPT-cloud” chemometrics toolbox [7]. An example of experimental Raman spectra of malignant melanoma after preprocessing is shown in figure 1.

![Figure 1. An example of in vivo Raman spectra of MM after preprocessing.](image)

The study involved 208 patients. Raman spectra of normal skin and skin neoplasm were registered for each patient. In total, we used 416 spectra: 208 of normal skin, 80 of BCC, 49 of MM, and 79 of PN. All in vivo studies are conducted on patients older than 18 years of age and with their consent. The studies were approved by the ethics committee of Samara State Medical University (Samara, Russia).

2.2. Analysis approach
The first step of the analysis was unmixing Raman spectra into their components by a multivariate curve resolution using alternating least squares (MCR-ALS) analysis. We used a protocol by Felten et al. [8]. The main idea of MCR-ALS is to decompose the Raman spectra matrix $D$ into smaller matrices $C$ and $S^T$:

$$D = C \cdot S^T + E,$$  

(1)
where $C$ represents the concentration profiles for each of the skin components, $S^T$ is the pure component spectra matrix, and $E$ is the error matrix.

After initial estimation is given for $C$, it is optimized iteratively using an alternative least squares (ALS) algorithm until convergence is reached [8].

In our study the Raman spectra of the skin have been unmixed into eight components. We have chosen this number of components for two reasons. Firstly, Feng et al. [9] reports that eight components captured the skin constituents as measured on in vivo human skin cancers. They are collagen, elastin, triolein, nucleus, keratin, ceramide, melanin, and water. Secondly, we experimentally confirmed that the classification model gives the best results when unmixing into exactly eight components.

It should be noted here that we apply the MCR-ALS analysis in such a way that we only know the matrix $D$. The matrix $S^T$ is initially unknown. Therefore, we do not know which real skin components correspond to the components we found.

The second step of the analysis was the construction of a binary classifier system for disease classification. We applied the partial least squares discriminant analysis (PLS-DA) to the data obtained in the first step, the matrix of concentrations $C$ from (1). We used the “TP3-cloud” chemometrics toolbox for that [7].

3. Results and discussions

We have obtained binary classification models for the following cases:

- normal skin vs. neoplasms
- normal skin vs. BCC
- normal skin vs. MM
- BCC vs. MM
- PN vs. MM
- benign neoplasms vs. malignant neoplasms

As a result the ROC curve was calculated and the optimal classification threshold was chosen for each case. The accuracy for each case classification model (that is, the proportion of true results, either true positive or true negative) and the corresponding sensitivity and specificity are presented in Table 1. As one can see from Table 1, the classification models “normal skin vs. MM” and “BCC vs. MM” have high classification accuracy, which are 0.8673 and 0.8372, respectively. In addition, the model “normal skin vs. neoplasms” has a fairly high accuracy (0.7212). This can be explained by the fact that melanin makes a large contribution to the Raman spectrum of melanoma. It is generally one of the most relevant skin components [9]. If there is a lot of melanin in the MM, then in the case of BCC and normal skin, the concentration of melanin is not high.

The model “PN vs. MM” has the lowest accuracy among other models. This can also be explained by the presence of melanin in neoplasms: the concentration of melanin is high both in MM and in PN [9].

| Table 1. An accuracy, sensitivity and specificity for different binary classification models. |
|------------------------------------------------------------------------------------------|
| Model                                    | Accuracy | Sensitivity | Specificity |
|-------------------------------------------|----------|-------------|-------------|
| Normal skin vs. neoplasms                 | 0.7212   | 0.6462      | 0.7981      |
| Normal skin vs. BCC                       | 0.6750   | 0.5125      | 0.8375      |
| Normal skin vs. MM                        | 0.8673   | 0.8571      | 0.8776      |
| BCC vs. MM                                | 0.8372   | 0.7959      | 0.8625      |
| PN vs. MM                                 | 0.6328   | 0.5714      | 0.6709      |
| Benign neoplasms vs. malignant neoplasms  | 0.6923   | 0.7442      | 0.6076      |
The models “normal skin vs. BCC” and “benign neoplasms vs. malignant neoplasms” also have low accuracy, which are 0.6750 and 0.6923, respectively. In the first case, the concentration of melanin is low in both normal skin and BCC, and in the second case, it is high in PN (benign neoplasm) and MM.

Figure 2. ROC curves for different binary classification models: a – normal skin vs. neoplasms, b – normal skin vs. BCC, c – normal skin vs. MM, d – BCC vs. MM, e – PN vs. MM, f – benign neoplasms vs. malignant neoplasms.
The ROC curves for different binary classification models are presented in figure 2. The higher the ROC curve the better the fit of the classification model. The area under the curve (AUC) can be used for this purpose. The closer AUC is to 1 the better the fit. As one can see from figure 2, model “normal skin vs. MM” shows the best fit (AUC = 0.93). The calculated values of 0.86 and 0.88 for “normal skin vs. neoplasms” and “BCC vs. MM” also show a pretty good fit. The models “normal skin vs. BCC” and “PN vs. MM”, conversely, show a weak fit.

Our findings suggest that our approach could also be useful for classification of skin neoplasms. However, the approach has limitations. So, apparently, the concentration of melanin in the sample has a great influence on the classification result. This makes it possible to classify successfully neoplasms with high and low melanin content, for example, MM and BCC. At the same time, the classification of neoplasms with approximately the same content of melanin (for example, PN and MM) becomes difficult. It is advisable to apply other methods or classification criteria for these purposes.

Acknowledgments
This study was supported by the RFBR Grant № 19-52-18001 Bolg_a.

References
[1] “Cancer,” World Health Organisation, 3 March 2021, https://www.who.int/news-room/fact-sheets/detail/cancer
[2] “2020 Melanoma Mortality Rates Decreasing Despite Ongoing Increase in Incidence,” Melanoma Research Alliance, 9 January 2020, https://www.curemelanoma.org/blog/article/2020-melanoma-mortality-rates-decreasing-despite-ongoing-increase-in-incidence-rates#:~:text=2020%20Melanoma%20Facts%20and%20Figures,4%2C610%20men%20and%202%2C240%20women.
[3] Popp J, Krafft C, Mayerhöfer T 2011 Optik & Photonik 6(4) 24 – 8
[4] Bratchenko I A, Artemev D N, Khristoforova Y A, Bratchenko L A, Myakinin O O, Moryatov A A, Orlov A E, Kozlov S V and Zakharov V P 2020 Raman Spectroscopy Techniques for Skin Cancer Detection and Diagnosis Multimodal Optical Diagnostics of Cancer ed V V Tuchin, J Popp, V Zakharov (Springer, Cham) pp 359 – 93
[5] Khristoforova Y A et al. 2020 Ultrafast Nonlinear Imaging and Spectroscopy VI - Proceeding of SPIE 11582
[6] Khristoforova Y A, Bratchenko I A, Myakinin O O, Artemev D N, Moryatov A A, Orlov A E, Kozlov S V and Zakharov V P 2019 Journal of biophotonics 12(4) p e201800400
[7] “TP3-cloud” chemometrics toolbox, https://tptcloud.com/#/
[8] Felten J, Hall H, Jaumot J, Tauler R, De Juan A and Gorzsás A 2015 Nature Protocols 10(2) 217
[9] Feng X, Moy A J, Nguyen H T M, Zhang J, Fox M C, Sebastian K R, Reichenberg J S, Markey M K and Tunnell J W 2017 Biomedical optics express 8(6) pp 2835 – 50