Mathematical Analysis of Raman Spectra Data Arrays Using Machine Learning Algorithms

Yana A. Byuchkova*, Andrey Y. Zyubin, Vladimir V. Rafalskiy, Ekaterina M. Moiseeva, and Ilia G. Samusev

Immanuel Kant Baltic Federal University, 14 Alexander Nevsky str., Kaliningrad 236016, Russian Federation

***e-mail:** 18377@mail.ru

Abstract. This paper is devoted to the application of mathematical methods of classification and differentiation of low-resolution spectral data arrays of Raman light scattering for complex biological compounds as human platelets. Spectral data arrays consisted of 1266 spectra from 4 groups of patients, totaling 152 people were analyzed. A random forest algorithm was used. Potential biomarkers of differences between patient groups were identified, on which the given algorithms were tested. Using the random forest algorithm for classification of spectra of healthy patients without therapy and patients with cardiovascular pathologies without therapy, we have achieved the accuracy of 83.4%. Classification of the healthy patients on and off therapy shows the accuracy of 76.26% and classification of the patients with cardiovascular pathologies shows 70% accuracy.

Keywords: spectroscopy; surface-enhanced Raman scattering; cardiovascular disease; machine learning; random forest algorithm.

Paper #8684 received 5 Mar 2023; revised manuscript received 13 Jun 2023; accepted for publication 16 Jun 2023; published online 29 Jun 2023. doi: 10.18287/JBPE23.09.020308.

1 Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide and in the Russian Federation. An estimated 17.9 million people died of cardiovascular disease in 2016, accounting for 31% of all deaths worldwide. Eighty five percent of these deaths were caused by heart attack and stroke [1]. Cardiovascular disease is the most frequent cause of hospitalization and disability among the population of the Russian Federation. At the same time, about 40% of people in Russia die at an active working age (25–64 years) [2].

The development of new methods for diagnosing and detecting the risks of such diseases can significantly increase the efficiency and accuracy of preventing these diseases. The application of machine learning algorithms for blood Raman spectra analysis is a relatively novel method for detecting cardiovascular diseases.

Research into the application of machine learning to the processing of Raman spectra for the diagnosis of cardiovascular diseases is at an early stage of development. For example, machine learning has been applied to early warning of heart attacks using surface-enhanced Raman spectroscopy [3].

A random forest algorithm was chosen to classify Raman spectra by patient groups and select the most significant spectral shifts [4–7]. This choice is due to the ability of this method to classify, the ability to select the most significant features and to deal with large amounts of data. It is also effective and relatively fast, which is important for the analysis of large amounts of data in a short time. The originality of the approach lies in the first application of machine learning to the processing of a spectral array of complex Raman light scattering spectra of human platelets.

The purpose of this paper is to develop a solution for an important problem of differentiation of Raman spectra [8–11] in patients with and without cardiovascular pathologies, detection of spectral markers of platelet changes in pathologies and due to medication.

This article describes the results of applying a machine learning algorithm to the processing of spectral data arrays for different groups of patients: healthy patients, patients with pathologies of cardiovascular disease, healthy patients receiving therapy, and patients with pathologies of cardiovascular disease receiving

This paper was presented on the Biophotonics Workshop of the XX All-Russian Youth Conference-Contest on Optics and Laser Physics, Dedicated to the 100th Anniversary of N.G. Basov, Samara, Russia, November 8–12, 2022.

therapy. The applicability of the random forest algorithm is shown.

2 Materials and Methods

2.1 SERS Experiment

For the platelet study using SERS, fresh venous blood samples were collected from healthy volunteers and patients with CVD in a vacuum tube containing EDTA (BD Vacutainer®). The samples were then centrifuged at 60 g for 15 min to separate platelet-rich plasma, and then it was centrifuged again with 60 g for 15 min to precipitate leukocytes and erythrocytes. In the last step, platelets were precipitated by centrifugation of the supernatant at 1500 g for 15 min. All centrifugation steps were performed at 4 °C.

Fresh venous blood samples for SERS were taken in a vacuum tube containing sodium citrate (Vacutainer 4.5 ml with 3.2%® sodium citrate). SERS spectra were obtained on a Centaur U HR Raman spectrometer (NanoScanTechnology LTD, Russia) using a Cobolt Samba diode-pumped solid-state laser with a photoexcitation wavelength of $\lambda = 532$ nm and a power per sample of 45 mW. The optical scheme included an Olympus BX 41 microscope with a 100X objective (NA 0.9). The spectrometer monochromator (ZAO Solar LS, Belarus) had a focal distance of 284 mm with a diffraction grating of 1200 gr/mm and was equipped with a IDus 401 CCD camera (Andor, UK). The resolution of camera was 1024 × 256 pixels. The spectrometer had a spectral resolution of 2.5 cm^{-1}. A 1 × 25 μm laser spot was positioned manually on the platelet mass. Rayleigh scattering was eliminated with a reflector filter. A drop of 5 μl of platelet-rich plasma was applied to a previously created titanium-based nanostructured surface, a detailed methodology of production of which is described in Ref. [12] substrate, dried for 5 min at room temperature, and then placed in a microscope rack. For each sample, spectra averaged three times from ten different drop locations were collected. The exposure time for the CCD camera was set as 70 s. Each time before the experiment, the spectrometer was calibrated for silicon by a static spectrum centered at a spectral shift frequency of 520.1 cm^{-1} for 1 s. After registration, the spectra were saved in .txt format and in a special format (.ngs) on a PC connected to the Raman spectrometer. Due to the possibility of plasmon resonance generation, roughened titanium surfaces with spherical gold nanoparticles were used to amplify the Raman signal. The plasmon absorption maxima were $\lambda = 530$ nm and $\lambda = 570$ nm for gold nanoparticles and the rough Ti surface, respectively. For such structures, the spectral signal amplification was recorded up to orders of 103 times.

The study focused on the Raman spectra of the following groups of patients:

- healthy without therapy (group 1),
- healthy on therapy (aspirin, clopidogrel) (group 2),
- patients with cardiovascular disease without therapy (group 3),
- patients with cardiovascular disease on therapy (group 4).

The total number of study participants was 152. Subjects were divided into 4 groups: healthy volunteers not receiving AT (group 1) and receiving acetylsalicylic acid (ASA), group 2; patients with cardiovascular disease (CVD) without AT (group 3) and receiving AT (group 4). Aggregometry and SERS spectra of platelets were performed in all subjects. An original optical sensor based on gold particle-modified nanostructured titanium surface was developed to obtain SERS spectra of platelets.

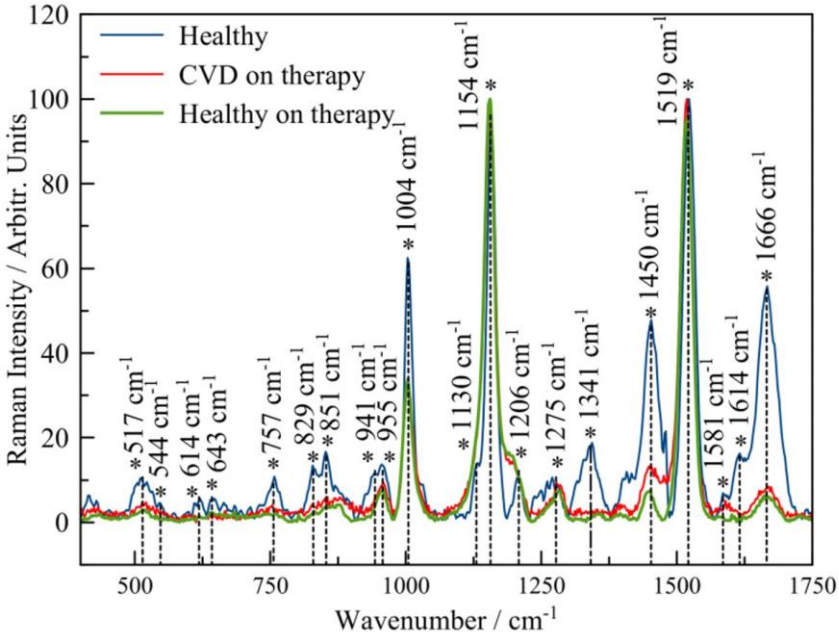

Fig. 1 Example of Raman spectra for healthy patients (blue line), patients with CVD on therapy (red line), group of healthy patients on therapy (green line) [12].

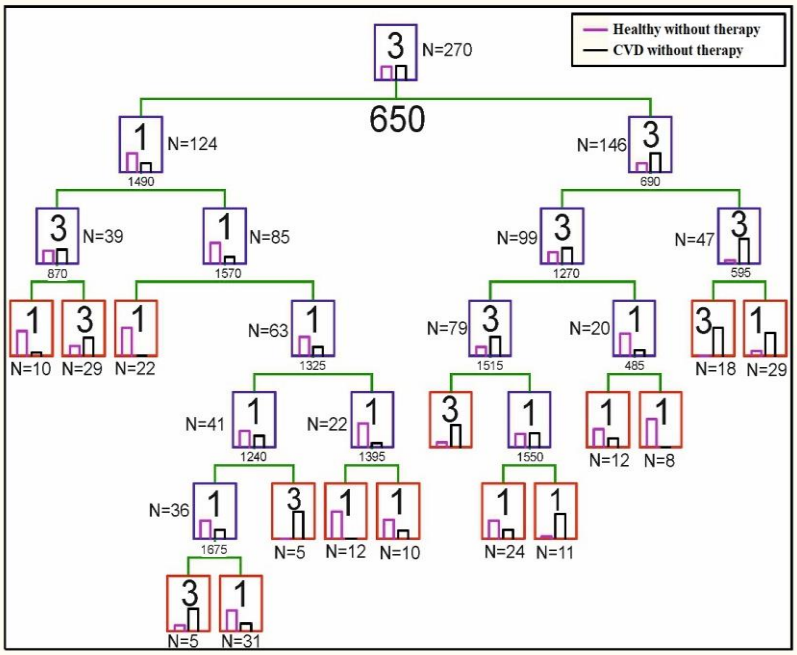

Fig. 2 Decision tree in the classification of spectra by groups of healthy patients without therapy (group 1) and patients with cardiovascular disease pathologies without therapy (group 3).

Spectrum data were read from the instrument and entered into .txt and .csv files. For the following study, spectral data were brought to uniformity by classifying a single frequency grid using the Parcer program. Uniform group tables were created and a unified frequency grid of 5 cm^{-1} spacing was generated. Thus, the developed program automatically converted files and formed an array of data. This program was written in C++ and arranged the spectrum by grid cells in the range from 400 cm^{-1} to 1800 cm^{-1} in 5 cm^{-1} increments. Thus, all spectral fluctuations were correlated with the designated grid.

2.2 Spectral Data Processing with Machine Learning

The Statistica 13 Random Forest module, which is an implementation of the random forest classifiers developed by Breiman [13], whose algorithm is also applicable to regression tasks, was chosen to process spectral data.

To implement the random forest classification tasks, a group column was added to the data table, which included the name of the patient group of the selected observation. The dependent variable was the group distribution of patients according to their condition and medication intake. And spectral shift values between 400 cm^{-1} and 1800 cm^{-1} were taken as continuous predictors. Next, a test sample of spectral observations was allocated using the additional option of selecting the proportion of test and training samples. In our study, the test sample represents 30% of all observations. The random forest method in Statistica 13 defines a boundary function that measures the extent to which the mean number of votes for the correct class exceeds the mean number of votes for any other class present in the dependent variable. This measure provides us not only with a convenient way to make predictions, but also a way to relate the validity score to those predictions. The accuracy of the classifier was determined by the ratio of the number of correct responses to the total number of responses.

3 Results and Discussions

An array of spectral data from four groups of patients was accumulated based on the results of spectral imaging. An example of the obtained spectra is shown in Fig. 1.

Table 1 Classification matrix for separating healthy patients without therapy from CVD patients without therapy.

	Predicted healthy without therapy	**Predicted CVDs without therapy**
True healthy without therapy	37.5%	10.07%
True CVD without therapy	6.53%	45.9%

Note: The number of correctly classified spectral data by patient group is highlighted in green. The number of misclassified spectral data is highlighted in red.

Table 2 The most important spectral shifts and their interpretation for the classification of healthy patients without therapy from CVD patients without therapy.

Spectral shift, cm^{-1}	Significance of the trait, relative units	Interpretation	Reference
1305	0.00803	$\tau(CH_2)$ (lipids)	[14]
1485	0.00782	1485 planar variations (adenine/guanine)	[15]
1155	0.00678	C–C valence oscillations (v) in proteins	[16–17]
1395	0.00671	Amid	[18]
890	0.00621	CH_3 bend in Tricaprylin	[14]

Table 3 Classification matrix to separate healthy patients without therapy from healthy patients on therapy.

	Predicted healthy without therapy	Predicted CVDs without therapy
True healthy without therapy	69.38%	11.88%
True healthy on the therapy	11.88%	6.88%

Note: The number of correctly classified spectral data by patient group is highlighted in green. The number of misclassified spectral data is highlighted in red.

Table 4 The most important spectral shifts and their interpretation for the classification of healthy patients without and on therapy.

Spectral shift, cm^{-1}	Significance of the trait, relative units	Interpretation	Reference
1040	0.00714	Unsaturated fatty acids	[14]
1035	0.00707	β(CH) in lipids	[14]
1685	0.00607	Protein beta sheet and polyproline helix	[19]
980	0.00571	β(CH) in lipids	[14]
1165	0.005	ν(C-C) in lipids	[14]
1305	0.005	$\tau(CH_2)$, δ(=CH) in lipids	[14]

3.1 Spectral Data Processing with Machine Learning

The data were classified into patient groups using a random forest algorithm. Let us first consider differentiation of spectra by groups of healthy patients without therapy (group 1) and patients with cardiovascular pathologies without therapy (group 3). Fig. 2 shows one of the decision trees, which shows the most informative spectral lines.

Most of the data were correctly identified using the random forest algorithm. The correctness of the algorithm for classifying the observations into groups of healthy patients without therapy and patients with cardiovascular pathology without therapy was 83.4%, as shown in Table 1.

In carrying out this classification, the most significant spectral shifts in separating the observations into groups were highlighted, as presented in Table 2.

Table 5 Classification matrix for separating non-therapy CVD patients from therapy naive CVD patients.

	Predicted CVDs without therapy	**Predicted CVDs on therapy**
True CVD without therapy	59.5%	3.0%
True CVD on therapy	27.0%	10.5%

Note: The number of correctly classified spectral data by patient group is highlighted in green. The number of misclassified spectral data is highlighted in red.

Table 6 The most important spectral shifts and their interpretation for the classification of CVD patients without and on therapy.

Spectral shift, cm^{-1}	**Significance of the trait, relative units**	**Interpretation**	**Reference**
1565	0.00662	C–N stretching	[20]
1425	0.00655	The band at ~1420 cm^{-1} observed for saturated and Z-unsaturated FAs is hardly seen in the spectra of the unsaturated compounds and presumably accounts for the shoulder at approximately 1422 cm^{-1}	[21]
1055	0.00615	Lipid hydrocarbon chain	[14]
1440	0.00602	CH_2 bend (lipids)	[16–17]
1385	0.00582	Hydrogen bonding in protein	[22]
1125	0.00576	V($C_ß$-methyl)	[23]

Also, a random forest algorithm was used to classify healthy patients on and off therapy with an accuracy of 76.26%, as can be seen in Table 3.

The most significant spectral shifts and their interpretation for the classification of healthy patients without and on therapy were highlighted, shown in Table 4.

When the patients with cardiovascular pathologies on and off therapy were classified, 70% accuracy was achieved, as shown in Table 5.

Table 6 highlights the most important spectral shifts and their interpretation for the classification of patients with cardiovascular pathologies without and on therapy.

4 Conclusions

As a result of this research, the initial results on the development of machine learning methods for the differentiation of Raman spectra in patients with and without cardiovascular disease pathologies were demonstrated.

Approaches were applied to process spectral data arrays consisting of 1266 spectra for different groups of patients: healthy patients, patients with pathologies of cardiovascular diseases, healthy patients receiving therapy and patients with pathologies of cardiovascular diseases receiving therapy. The applicability of the random forest algorithm was shown. Potential biomarkers of differences between patient groups on which the given algorithms were tested were identified. The achieved accuracy of classification using the random forest algorithm of spectra across the groups of healthy patients without therapy and patients with cardiovascular pathologies without therapy was 83.4%. Classification of the healthy patients on and off therapy shows the accuracy of 76.26% and classification of the patients with cardiovascular pathologies shows 70% accuracy.

Acknowledgements

This research was supported from the Russian Federal Academic Leadership Program Priority 2030 at the Immanuel Kant Baltic Federal University.

Disclosures

The authors have no conflict of interest to disclose.

References

1. World Health Organization, Cardiovascular Diseases (CVDs), (accessed 6 May 2023). [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)].
2. N. N. Pribylova, O. A. Osipova, M. A. Vlasenko, O. A. Vlasenko, and A. Y. Chetverikova, "Diagnostic aspects of definition of operational damage of a myocardium at a coronary revascularization," Challenges in Modern Medicine 18(10), 17–23 (2012). [in Russian]
3. Z. Liu, D. Meng,G. Su, P. Hu, B. Song, Y. Wang, W. Junhan,Y. Hao, Y. Tianyi, C. Buyun, O. Tse-Hsien, H. Sushmit, M. Matthew, L. Fanxin, and W. Wu, "Ultrafast Early Warning of Heart Attacks through Plasmon-Enhanced Raman Spectroscopy using Collapsible Nanofingers and Machine Learning," Small 19(2), 2204719 (2023).
4. F. B. de Santana, W. B. Neto, and R. J. Poppi, "Random forest as one-class classifier and infrared spectroscopy for food adulteration detection," Food Chemistry 293, 323–332 (2019).
5. B. P. Lovatti, M. H. Nascimento, K. P. Rainha, E. C. Oliveira, Á. C. Neto, E. V. Castro, and P. R. Filgueiras, "Different strategies for the use of random forest in NMR spectra," Journal of Chemometrics 34(12), e3231 (2020).
6. A. Wójtowicz, J. Piekarczyk, B. Czernecki, and H. Ratajkiewicz, "A random forest model for the classification of wheat and rye leaf rust symptoms based on pure spectra at leaf scale," Journal of Photochemistry and Photobiology B: Biology 223, 112278 (2021).
7. S. Khan, R. Ullah, A. Khan, A. Sohail, N. Wahab, M. Bilal, and M. Ahmed,"Random forest-based evaluation of Raman spectroscopy for dengue fever analysis," Applied Spectroscopy 71(9), 2111–2117 (2017).
8. G. Li, D. Wang, J. Zhao, M. Zhou, K. Wang, S. Wu, and L. Lin, "Improve the precision of platelet spectrum quantitative analysis based on "M+N" theory," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 264, 120291 (2022).
9. L. Hu, C. Yin, S. Ma, and Z. Liu, "Rapid detection of three quality parameters and classification of wine based on Vis-NIR spectroscopy with wavelength selection by ACO and CARS algorithms," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 205, 574–581 (2018).
10. T. Chen, Q. Chang, J. G. P. W. Clevers, and L. Kooistra, "Rapid identification of soil cadmium pollution risk at regional scale based on visible and near-infrared spectroscopy," Environmental Pollution 206, 217–226 (2015).
11. R. K. Douglas, S. Nawar, M. C. Alamar, A. M. Mouazen, and F. Coulon, "Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques" Science of the Total Environment 616, 147–155 (2018).
12. A. Zyubin, V. Rafalskiy, A. Tcibulnikova, E. Moiseeva, K. Matveeva, A. Tsapkova, I. Lyatun, P. Medvedskaya, I. Samusev, and M. Demin, "Surface-enhanced Raman spectroscopy for antiplatelet therapy effectiveness assessment," Laser Physics Letters 17(4), 045601 (2020).
13. L. Breiman, "Random forests,"Machine Learning 45, 5–32 (2001).
14. K. Czamara, K. Majzner, M. Z. Pacia, K. Kochan, A. Kaczor, and M. Baranska, "Raman spectroscopy of lipids: a review," Journal of Raman Spectroscopy 46(1), 4–20 (2015).
15. A. J. Hobro, M. Rouhi, E. W. Blanch, and G. L. Conn, "Raman and Raman optical activity (ROA) analysis of RNA structural motifs in Domain I of the EMCV IRES," Nucleic Acids Research 35(4), 1169–1177 (2007).
16. D. Garc´ıa-Rubio, B. de la Mora, I. Badillo-Ram´ırez, D. Cerecedo, J. Saniger, J. Ben´ıtez-Ben´ıtez, and M. Villagr´an-Muniz, "Analysis of platelets in hypertensive and normotensive individuals using Raman and Fourier transform infrared-attenuated total reflectance spectroscopies," Journal of Raman Spectroscopy 50(4), 509–521 (2019).
17. J. Depciuch, E. Kaznowska, I. Zawlik, R. Wojnarowska, M. Cholewa, P. Heraud, and J. Cebulski, "Application of Raman spectroscopy and infrared spectroscopy in the identification of breast cancer," Applied Spectroscopy 70(2), 251–263 (2016).
18. E. M. Jones, G. Balakrishnan, T. C. Squier, and T. G. Spiro, "Distinguishing unfolding and functional conformational transitions of calmodulin using ultraviolet resonance R aman spectroscopy," Protein Science 23(8), 1094–1101 (2014).
19. N. C. Maiti, M. M. Apetri, M. G. Zagorski, P. R. Carey, and V. E. Anderson, "Raman spectroscopic characterization of secondary structure in natively unfolded proteins: α-synuclein," Journal of the American Chemical Society 126(8), 2399–2408 (2004).
20. P. Schellenberg, E. Johnson, A. P. Esposito, P. J. Reid, and W. W. Parson, "Resonance Raman scattering by the green fluorescent protein and an analogue of its chromophore," The Journal of Physical Chemistry B 105(22), 5316–5322 (2001).
21. W. Curatolo, S. P. Verma, J. D. Sakura, D. M. Small, G. G. Shipley, and D. F. H. Wallach, "Structural effects of myelin proteolipid apoprotein on phospholipids: a Raman spectroscopic study," Biochemistry 17(9), 1802–1807 (1978).
22. H. Takeuchi, "Raman structural markers of tryptophan and histidine side chains in proteins," Biopolymers: Original Research on Biomolecules 72(5), 305–317 (2003).
23. B. R. Wood, P. Caspers, G. J. Puppels, S. Pandiancherri, and D. McNaughton, "Resonance Raman spectroscopy of red blood cells using near-infrared laser excitation," Analytical and Bioanalytical Chemistry 387, 1691–1703 (2007).