The use of data mixing to increase the size of the dataset for colon cancer diagnosis using diffuse reflectance spectroscopy and machine learning

Valentin Kupriyanov[1,2] , Maria R. Pinheiro[3], Sónia D. Carvalho[4,5], Isa C. Carneiro[4,6], Rui M. Henrique[4,7] ,Valery V. Tuchin[2,8,9], Luís M. Oliveira [3,10], Marine Amouroux[1], Yury Kistenev[2] and Walter Blondel[1]

[1] Université de Lorraine, CNRS, CRAN UMR 7039, Vandoeuvre-Lès-Nancy, France
[2]Laboratory of Laser Molecular Imaging and Machine Learning, Tomsk State University, Tomsk, Russia
[3]Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto, Portugal
[4]Department of Pathology and Cancer Biology and Epigenetics Group, Portuguese Oncology Institute of Porto, Porto, Portugal
[5]Department of Pathology, Santa Luzia Hospital (ULSAM), Viana do Castelo, Portugal
[6]Department of Pathological, Cytological and Thanatological Anatomy, Polytechnic of Porto – School of Health (ESS), Porto, Portugal
[7]Department of Pathology and Molecular Immunology, Porto University – Institute of Biomedical Sciences Abel Salazar, Porto, Portugal
[8]Science Medical Center, Saratov State University, Saratov, Russian Federation
[9]A. N. Bach Institute of Biochemistry, RC "Biotechnology of the Russian Academy of Sciences," Moscow, Russian Federation
[10]Physics Department, Polytechnic of Porto – School of Engineering (ISEP), Porto, Portugal
valentin.kupriyanov@univ-lorraine.fr

The research was carried out with the support of a grant under Vernadski international joint PhD program (2021-2024) provided by the French embassy in the Russian Federation
The research was carried out with the support of a grant under the Decree of the Government of the Russian Federation No. 220 of 09 April 2010 (Agreement No. 075-15-2021-615 of 04 June 2021)
This research was supported partly by the French PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE.

The use of optical methods in combination with machine learning techniques to diagnose various pathologies and diseases is a very promising branch of science. However, in order to obtain accurate and reliable machine learning models, a large number of samples is required, which is often impossible even in the case of medical research. This problem can be partially solved by using data generation techniques, the simplest and most accessible of which is mixing. This study presents the results of using different mixing-based data generation strategies to increase the size of a dataset of diffuse reflectance spectra measured on XX? healthy and cancerous colon tissue *ex vivo* samples in order to train a classification model. The experimental set-up consisted of an integrating sphere coupled to a broadband deuterium-halogen light source and to a XXX spectrometer to acquire reflected intensity spectra in the range from 230 nm to 900 nm. The mixing of the data was implemented in three ways: mixing with randomly chosen weights for the original spectra, mixing using one randomly chosen spectrum as a basis one and the rest as auxiliaries, and mixing using multiple randomly chosen spectra with equal weights. This contribution presents a comparison of the performance of these strategies on the results of classification of diffuse reflectance spectra of healthy and cancerous colon tissues.

Keywords: colorectal cancer, diffuse reflectance spectroscopy, machine learning, mixing.