Supplemental Document

Biomedical Optics EXPRESS

Machine learning-based LIBS spectrum analysis of human blood plasma allows ovarian cancer diagnosis: supplement

ZENGQI YUE,[1] CHEN SUN,[1] FENGYE CHEN,[1] YUQING ZHANG,[1] WEIJIE XU,[1] SAHAR SHABBIR,[1] LONG ZOU,[1] WEIGUO LU,[2] WEI WANG,[3] ZHENWEI XIE,[2] LANYUN ZHOU,[2] YAN LU,[2,4] AND JIN YU[1,5]

[1] *School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China*
[2] *Center for Uterine Cancer Diagnosis & Therapy Research of Zhejiang Province, Women's Reproductive Health Key Laboratory of Zhejiang Province, and Department of Gynecologic Oncology, Women's Hospital and Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou 310011, China*
[3] *Department of Clinical Laboratory, Tongde Hospital of Zhejiang Province, Hangzhou 310012, China*
[4] *yanlu76@zju.edu.cn*
[5] *jin.yu@sjtu.edu.cn*

Supplement DOI: https://doi.org/10.6084/m9.figshare.14329094

Parent Article DOI: https://doi.org/10.1364/BOE.421961

Machine learning-based LIBS spectrum analysis of human blood plasma allows ovarian cancer diagnosis: supplemental document

The classification model training process in this work was based on our previous work initially developed for quantitative analysis with LIBS spectra from soil samples with a back-propagation neural network (BPNN) [1]. In the present work, the method was adapted to the case of classification and identification of a collection of samples. The used neural network had 3 layers, with an input layer of 100 neurons corresponding to the 100 standardized selected features of each pretreated training spectrum, a hidden layer of 50 neurons, and an output layer of 3 neurons corresponding to the 3 output case-types. A 5-fold cross-validation optimization procedure was employed for neural network training. The implementation was applied to the ensemble of training spectra which is represented in Fig. S1, where a pretreated spectrum S_{ijk} is the k^{th} replicate of the j^{th} sample in the i^{th} case-type, and each pretreated spectrum contained 100 standardized selected spectral features.

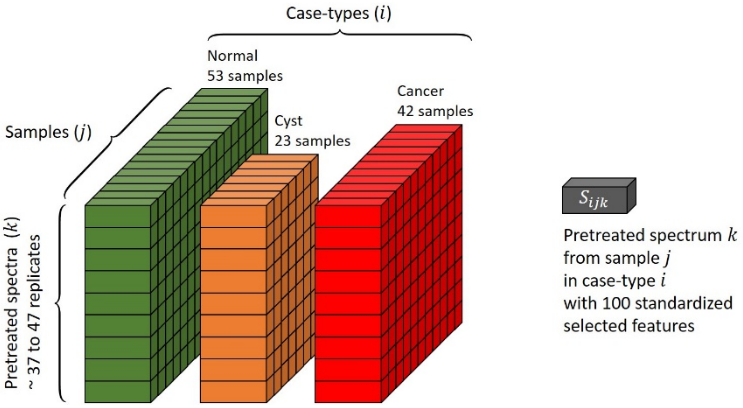

Fig. S1. Structure of the model training data set. An individual pretreated spectrum S_{ijk} is the k^{th} replicate of the j^{th} sample in the i^{th} case-type, with 100 standardized selected spectral features.

Since all the replicate pretreated spectra were statistically equivalent, the index k of a pretreated spectrum S_{ijk}, was in fact a dummy one. A data configuration could be thus obtained with a randomly arrangement of the replicates of each sample of the training sample set. Given a such data configuration, the replicates of each sample were divided into 5 subsets containing an equal (or almost equal) number of pretreated spectra $\{S_{ij\{k_1\}}\}, \{S_{ij\{k_2\}}\} \dots, \{S_{ij\{k_5\}}\}$. The subsets of the different samples were then associated in such way that the training data set was divided into 5 subsets, containing each an equal (or almost equal) number of pretreated spectra from the 3 case-types of normal, cyst and cancer, $\{S_{\{k_1\}}\}, \{S_{\{k_2\}}\} \dots, \{S_{\{k_5\}}\}$. A 5-fold iteration of cross-validation training by optimization with gradient descent then started with the first subset $\{S_{\{k_1\}}\}$ as the test spectra, while the ensemble of the rest 4 subsets as the training spectra. The first iteration generated a model (1) which was tested with the subset $\{S_{\{k_1\}}\}$, leading to an ensemble of identifications (1) for all the training samples. An identification of a sample among the 3 case-types of normal, cyst and cancer, was decided according to the majority of the individual identifications with the test spectra of the sample. A second iteration repeated the above process by using the second subset $\{S_{\{k_2\}}\}$ as the test spectra, while the ensemble of the rest 4 subsets as the training spectra, leading to an ensemble of identifications (2) for all the training samples. In the end of the 5 iterations, all the individual training spectra participated once as a validation spectrum. And the 5 ensemble of identifications (1) to (5) were generated with the 5 trained classification models. An ensemble of definitive identifications was assigned to all the training samples according to the majority of the 5 cross-validation identifications of a sample. The calibration performance of the trained models was then assessed by a comparison between the models-assigned case-type of each sample and their label value, and presented in the confusion matrix of the training samples.

References

1. C. Sun, Y. Tian, L. Gao, Y. S. Niu, T. L. Zhang, H. L, Y. Q. Zhang, Z. Q. Yue, N. D. Gilon, J. Yu, “Machine learning allows calibration models to predict trace element concentration in soils with generalized LIBS spectra,” Sci. Rep. **9**(1), 1–18 (2019).