Cloud technologies and machine learning in malignant tumors identification via Raman spectroscopy

Alexey S. Kovtunenko[1], Azat R. Bilyalov[2], Valentin N. Pavlov[2]

[1]*Ufa State Aviation Technical University, Ufa, Russia,*

[2] *Bashkir State Medical University, Ufa, Russia*

askovtunenko@ugatu.su*, azat.bilyalov@gmail.com*, *pavlov@bashgmu.ru

***Abstract* — the technical aspects of solving the problem of malignant tumors diagnostic using machine learning are considered. An algorithm of Raman spectra classification using machine learning is offered. This allows the differentiation of malignant and benign tumor tissues using only Raman spectroscopy. Also the special architecture of cloud system is offered, which allows to collect sample data from different medical institutions and research centres, store them using distributed storage technology. Using of the offered system allows to machine learning researchers apply the results of their investigations to medical diagnostic.**

***Keywords* – machine learning, Raman spetroscopy, distributed storage, malignant tumor, wavelet transformation.**

I. INTRODUCTION

Early diagnosis of cancer is critical for timely, effective and, ultimately, successful treatment. Changes in the structure and concentration of the main biochemical substances of cells and tissues begin long before the onset of clinical symptoms of a malignant tumor. In this regard, spectroscopy, which allows detecting changes in chemical bonds in molecules, is a potential tool for early diagnosis of tumors. As a method of molecular spectroscopy, Raman spectroscopy can detect cancer-induced changes in the molecular structure and composition of tissue.

Raman spectroscopy is an optical method that is currently being considered to characterise many diseases, including in-vivo applications that demonstrate the differences between benign and malignant tumor tissues. In this regard, the problem of developing a high-precision method for recording spectrograms, leading to the development of accurate models that can subsequently be used with in vivo tools to assess the tissue condition and determine tumor boundaries is urgent.

Interest in biological tissue spectroscopy is growing rapidly, as both clinical and non-clinical researchers have recognized that vibrational spectroscopic methods, infrared (IR) and Raman spectroscopic methods can potentially become non-invasive tools for diagnosing diseases. However, there is a significant gap in the development of spectrogram analysis methods, since it seems that the details of the characteristic peak frequencies and their definitions that can be attributed to specific functional chemical groups present in biological tissues, are not fully investigated. In addition, there is currently no single source that takes into account both IR and combinatorial spectroscopic studies of biological tissues, since researchers must rely on a number of research sources, and in most cases, the interpretation of spectral data varies significantly.

The purpose of this research is to develop a methodology for the analysis of Raman spectrograms using intelligent data processing methods. A new method of Raman spectra processing and interpreting presented in the article. Experimental verification includes collecting of Raman spectra of normal and malignant tissues samples, storing them in a special database and using them (after specific pre-processing) as a training sample for artificial neural network (ANN).

II. RAMAN SPECTROSCOPY IN CANCER DIAGNOSTICS

Raman spectroscopy of a biological object is a very non-trivial problem and requires a special approach.

Huang et al. provided information for interpreting peak intensities of biological samples [1], [2]. Proteins, lipids, nucleic acids and polysaccharides have distinct and detectable peaks in the Raman spectra. Some types of tumors have higher amounts of proteins, lower amounts of phospholipids and wider and higher peaks of nucleic acids. However, there are disparities between the peaks of different tissues.

Thus, prostatic tissue was investigated using vibrational spectroscopic methods. E. Gazi et al. FTIR microspectroscopy was used to differentiate samples of benign and cancerous prostate tissue implanted with paraffin. They also successfully differentiated prostate cancer cell lines obtained from various metastatic sites using FTIR spectroscopy. It was found that the ratio of the peak areas of 1030 and 1080 cm-1 (respectively, to the fluctuations in glycogen and phosphate) indicates a potential method for the differentiation of benign from malignant cells. It should be noted that the tissues were analyzed after installation on the BaF2 plate and subsequent wax removal [3].

Paluszkiewicz, C., Kwiatek, W.M. investigated human prostate tissues using FTIR microspectroscopy and synchrotron-induced x-ray emission (SRIXE) methods. Tissue samples were also analysed by a histopathologist. In this study, differences between the cancerous and indistinct parts of the analysed tissues were observed for both methods [4].

978-1-7281-5764-1/19/$31.00 ©2019 IEEE

There are several key points in determining the corresponding functional groups of each peak. These points can play an outstanding role in the process of peak characteristic analysis, and they are of exceptional importance in the process of a clear understanding of the spectral methods used in research. There is an increasing need to identify key spectral peaks and their proper assignment to chemical structures. Therefore, it is extremely important to have a reliable spectral database that is widely available to researchers working with vibration spectroscopic methods.

In spectroscopic studies, the precise determination of peaks can have a significant impact on the reliability of the results. Spreading in the literature, it becomes apparent that most scientists have mainly used previously published articles to determine the data obtained from the collected spectra. However, without a reliable and detailed database that can cover most of the known peaks in the spectral range, it would be a laborious task to find the value of different peak intensities.

III. Processing and analysis of the spectrograms

Signal processing technologies for automatic recognition, in general, include the following steps: preliminary signal processing, presenting the signal in the required form, features identification, error correction, selection and classification of features. Signal processing technologies such as basic component analysis (PCA), linear discriminant analysis (LDA), artificial neural network (ANN), support vector machine (SVM), logistic regression analysis (LRA) or a combination of these were used to analyze the Raman spectra of biological samples.

In many practical problems, researchers are mainly interested in features that exhibit the greatest variability (i.e., scatter, dispersion) during the transition from one studied object to another, while such features are often impossible to observe directly on the objects. That is, the task of reducing the dimensionality of multidimensional space appears, which consists in the expressing of a large number of initial factors directly measured at the objects through a smaller number of more capacious, most informative, but not directly observable internal characteristics of objects. PCA typically reduces high spectral data by a few most significant components [5].

The SVM method is an algorithm based on controlled learning, non-probabilistic models with appropriate learning algorithms for the purpose of regression analysis and classification, for the analysis and differentiation of data or models [6].

ANN is a computational algorithm that simulates the functioning of natural neural tissue. Harris et al. investigated the possibility of using ANN to distinguish between cancerous and normal cells by their Raman spectra [7]. Natural biological cells are a complex mixture of molecules (proteins, nucleic acids, lipids and sugars) that produce Raman spectra with a complex background. Therefore, preliminary work was carried out with well-characterized cultured cells under standardized laboratory conditions and Raman spectroscopy, to first understand the obtained spectra. ANN was then used to classify cancer cells from normal cells, reporting a 95% of specificity and 92% of sensitivity.

In regression analysis, logistic regression is estimating of parameters of a binary logistic model. Mathematically, the binary logistic model has a dependent variable with two possible values, such as malignant/benign, which is represented by an indicator variable, where two values are labeled "0" and "1". In a logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); each independent variable can be a binary variable or a continuous variable. The preliminary study compared the spectra of saliva samples from 45 healthy and 19 patients with lung cancer. An independent T-test sample was performed in the spectra. Distinctive peaks have been identified as biomarkers for the treatment of lung cancer. Discrimination by the LRA achieved an accuracy of 96.9%.

PCA-LDA processing methods are the most popular (45%), followed by PCA-ANN (33%) and SVM (22%). It should be noted that PCA-SVM gives the highest accuracy (99.9%), followed by PCA-ANN (98%) and LRA (97%). On the other hand, the performance of PCA-LDA and PCA-SVM is optimal both in sensitivity and specificity.

IV. Proposed methods and algorithms for Raman spectra processing and recognizing

In the article we propose a new approach to Raman spectra processing and recognizing. It consists of four main stages.

In the first stage, the raw spectrum, just obtained from spectroscope, is subjecting to filtering to get rid of possible digitisation errors. To reduce the influence of background fluorescence to further calculations, the baseline of the spectrum should also be corrected.

In the second stage, after filtering and baseline correction, spectrum is subjected to a discrete wavelet transform with obtaining an initial digital pattern of the tissue sample, which consists both of scaling factors and offset coefficients.

In the third stage, we obtain the final pattern as a result of multiplying the original pattern by the transformation matrix, obtained previously by the PCA algorithm. Obviously, PCA works with a full set of patterns. Therefore, in the third stage we should have the database with full set of samples patterns.

By multiplying by the transformation matrix with all the patterns, we obtain a training sample for further use for multilayer perceptron (MLP) training.

In the fourth stage, we train MLP on the training set obtained in the previous stage and after we can use it to recognize the final patterns (spectra passed through the first three stages) of unknown tissues samples. A well-trained MLP returns 1 if the tissue seems like benign and 0 otherwise.

V. Cloud architecture for collecting and storing of sample data

Cloud architecture [8] for storing anonymous results of diagnostic and lab tests, such are DICOM files from MRT and CT, data of different spectroscopic survey, digitized

cardiograms and encephalograms etc., includes the following set of computational nodes (Figure 1.).

- Client devices – sources and consumers of data. It could be diagnostic or lab equipment as well as researchers personal workstations (smartphones, laptops).
- Institutional (laboratory, personal) server – an element of distributed cloud software. It supports communication with client devices according to predefined access policy [9].

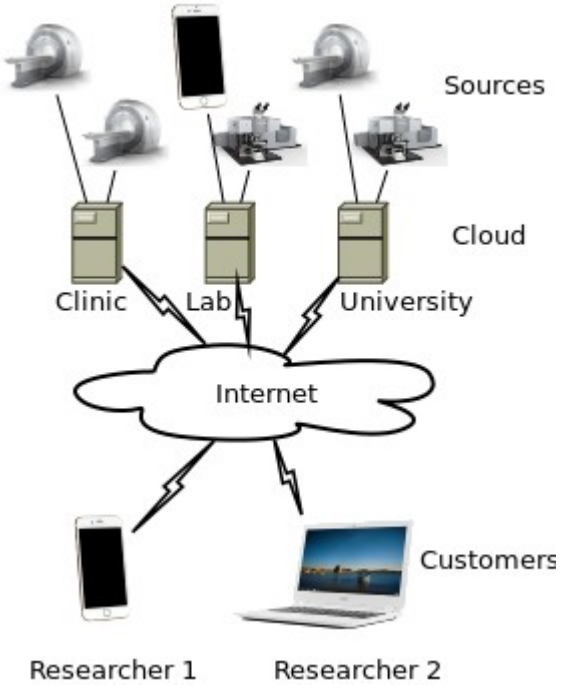


Figure 1. Architecture of cloud for distributed collecting and storing of anonymous medical data

Offered architecture is implemented in the form of distributed Nextcloud storage with additional software agents deployed on the workstations of medical and lab equipment [10]. The developed architecture includes special file formats that allow to store raw data (spectra, images), all intermediate data (filtered, corrected) and share the results of research.

VI. Experiment

Tumor and normal tissues obtained after surgical removal of malignant and benign tumors of patients of the Department of Urology and Oncology of the BSMU Clinic were used in the study.

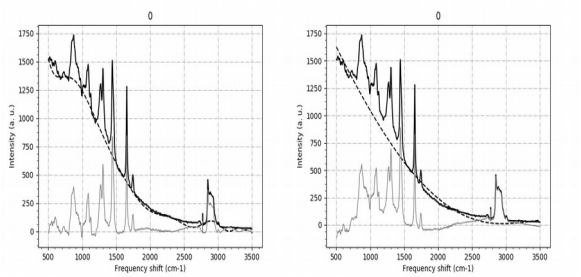
The section of postoperative preparation was carried out in the operating department, and its marking was in accordance with the standards of the BSMU Clinic. Tissue samples were wiped with a sterile cloth and transferred to 0.9% saline. Samples were delivered to the optical spectroscopy laboratory within 2 hours after the end of the operation. To obtain Raman spectra, we used the Horiba XploRA plus apparatus, Model BX 41 TF (Horiba, Ltd., Japan). A laser with a wavelength of 785 nm and a power of up to 100 mW was used to study biological tissues [11].

The data from the spectrograph were processed using the LabSpec v.6.4.4 software package. This software package performed graphical data analysis and conversion of the measurements into SPC formats – binary file format for storing spectroscopy data and CSV format for further processing using the developed analysis algorithm. Subsequent graphical analysis was performed using Spectragryph V software. 1.2.8.

Pre-processing, learning, and recognition algorithms were implemented using Python and the following libraries:

- scipy,
- peakutils,
- pywt,
- tensorflow,
- matplotlib.

The spectrum was preliminarily processed by the baseline adjustment method by iterative polynomial approximation (Figure 2.). This is one of the most commonly used baseline correction in spectroscopy. Figure 1a shows the use of a third-order approximation, and Figure 1b shows a sixth-order approximation. It can be seen that the second option gives a more accurate result.



(a) (b)

Figure 2. Spectrum baseline correction

The initial spectrum-based pattern was constructed using the wavelet transform. To identify the features of the spectrum, in which we will consider high peaks and eliminate noise in the experimental data, we used the expansion of the corrected spectrum into a bounded functional series in terms of the fourth-order Daubechies wavelets. This decomposition allows a multi-level analysis of the frequency spectrum abstracting from the physical meaning of the peaks in the spectrum [12].

A graphical representation of the spectrogram expansion using Daubechies wavelets after baseline adjusting is shown on Figure 3. Low-frequency components (located in the lower part of the picture) have significant differences, while in the middle part of the spectrum, repeating patterns can be observed. Thus, it can be concluded that it is theoretically possible to classify different tissues on the basis of a deeper analysis of the frequency decomposition of their combinational spectra. The result of the decomposition of each spectrum is a set of decomposition coefficients, the totality of which is an image in a multidimensional attribute space.

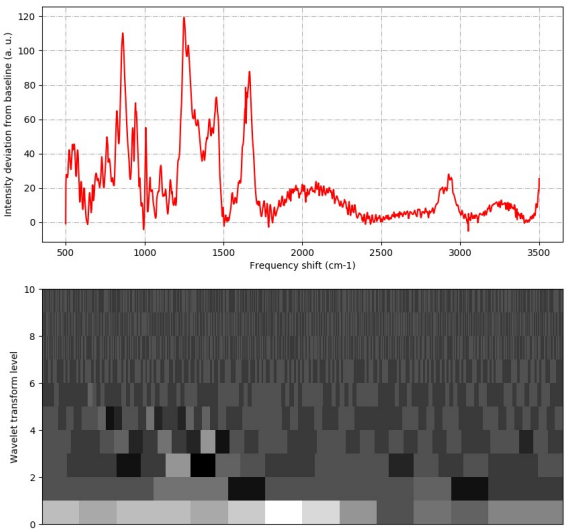


Figure 3. A graphical representation of the wavelet transform of the Raman spectrograms.

To reduce the dimension of the feature space and identify the most characteristic abstract properties of the spectra, based on the correlation analysis of the entire sample, the main components were selected and a projecting linear transformation of images from the original feature space to the reduced one was developed. The PCA method consists of the following main steps. First, the PCA converts function input data into orthogonal space using an orthogonal linear transformation. The result is orthogonal components, known as principal components. Secondly, the main components are organized according to their dispersion.

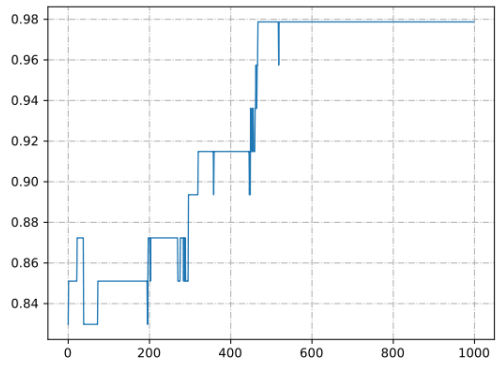


Figure 4. Training of the MLP

The difference is a measure of the variability of the sample distribution and is expressed as the mean square deviation of each sample from its average value.

To solve the problem of recognition in the reduced attribute space, modern computing technologies and artificial intelligence technologies based on machine learning methods and in-depth analysis of training samples were used. Figure 4. shows the learning outcomes for 10,000 epochs (50 epochs per point on the graph). At the same time, the recognition accuracy of the training data set was 97.5%.

VII. CONCLUSIONS

Main results.

- Intelligent malignant tissue recognition algorithm based on deep analysis of Raman spectra. The algorithm was implemented in Python using Deep Learning technologies. Reliability of identification of malignant tumor tissue ranged from 97.5% to 98%.
- Architecture of cloud system that allows to collect sample data from different medical institutions and research centres, stores them using distributed storage technology and also allows machine learning researchers to apply the results of their investigations to medical diagnostic. The offered architecture was implemented in the form of hardware and software complex. It is based on Nextcloud engine and allows integrating all diagnostic and lab equipment into unified system for collecting and storing anonymised medical data.

ACKNOWLEDGEMENT

The publication has been prepared with the support of the Department of Urology and Oncology of the Bashkir State Medical University Clinic and Institute of Oil Chemistry and Refining.

REFERENCES

[1] Huang, Z., McWilliams, A., Lui, M., McLean, DI, Lam, S., and Zeng, H. 2003. Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. Int. J. Cancer 107, 1047-1052.

[2] Huang, Z., Lui, H., McLean, DI, Korbelik, M., and Zeng, H. 2005. Raman spectroscopy in combination with background near-infrared autofluorescence enhances the in vivo assessment of malignant tissues. Photochem. Photobiol. 81, p. 1219-1226.

[3] Gazi, E., Dwyer, J., Gardner, P., Ghanbari-Siakhani, A., Wde, AP, Lockyer, NP, Vickerman, JC , Clarke, NW, Shanks, JH, Scott, LJ, Hart, CA, Brown, M., 2003. Applications of Fourier transform infrared microspectroscopy in studies of benign prostate and prostate cancer: A pilot study. J. Pathol. 201, 99 –108.

[4] Paluszkiewicz, C., and Kwiatek, WM, 2001. Analysis of human cancer prostate tissues using FTIR microscopy and SXIXE techniques. J. Mol. Structures 565–566, 329–334.

[5] McLachlan GJ. Discriminant analysis and statistical pattern recognition. New Jersey: John Wiley; 2004

[6] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20: 273-97.

[7] Harris A, Garg M, Yang X, Fisher S, Kirkham J, Smith D, et al. Raman spectroscopy and advanced mathematical modeling in the discrimination of human thyroid cell lines. Head Neck Oncol 2009; 1: 1-6.

[8] Alexey Kovtunenko, Marat Timirov, Azat Bilyalov. Multi-agent Approach to Computational Resource Allocation in Edge Computing // 918th International Conference, NEW2AN 2018, and 12th Conference, ruSMART 2018, St. Petersburg, Russia

[9] Kovtunenko, Alexey, Bilyalov, Azat, Valeev, Sagit. Distributed Streaming Data Processing in IoT Systems Using Multi-agent Software Architecture. 18th International Conference, NEW2AN 2018, and 11th Conference, ruSMART 2018, St. Petersburg, Russia

[10] Kovtunenko A. S., Valeev S. S., Maslennikov V. A. The multi-agent platform for the distributed real-time data processing. Natural and technical Sciences, No2(64), 2013

[11] Yakupov, R. R., Bilyalov, A. R., Kovtunenko, A. S. Use of machine learning for processing spectral characteristics of biological objects in diagnostic problems. Natural and technical Sciences, No10(124), 2018

[12] Bilyalov, A., Kovtunenko, A., Sysoeva, M. Diagnosis of human malignant tumors based on the raman spectra analysis using machine-learning. Information Technologies for Intelligent Decision Making Support ITIDS'2018 Proceedings of the 6 International Conference, Ufa, 2015.