A COMBINATION OF THE K-NEAREST NEIGHBOR ALGORITHM AND THE PRINCIPAL COMPONENT ANALYSIS FOR CLASSIFICATION AND SCREENING OF PATIENTS WITH CHRONIC LYMPHOCYTIC LEUKEMIA AND MULTIPLE MYELOMA.

R.V. Butyaev, D. A Chernyshev, L.V. Plotnikova, A.M. Polyanichko.

Saint Petersburg State University, Saint Petersburg

Chronic lymphocytic leukemia (CLL) is one of the most common oncohematological diseases. One of the important factors of successful therapy of such pathologies is their timely diagnosis. This work is aimed at developing screening approaches using infrared (IR) spectroscopy. Previously, we have shown the prospects of analyzing the IR spectrum of blood serum for the detection of multiple myeloma (MM) [1-4], however, it has not yet been possible to determine CLL in this way.

Our task was to evaluate the effectiveness of the method of analyzing the IR spectrum of blood serum in combination with machine learning methods - the principal component analysis (PCA) and the k–nearest neighbor's algorithm (kNN) for classifying blood serum samples with diagnosed CLL. Additionally, the effectiveness of the combination of these methods on MM samples will be evaluated.

Using the Nicolet (ThermoFisher) infrared Fourier spectrometer, the spectra of 8 blood serum samples from CLL patients and 20 healthy donors were taken. Measurements were carried out in KBr tablets in the range (4000-800) cm^{-1} with a resolution of 2 cm^{-1} and averaging over 128 accumulations. The NIPALS algorithm (nonlinear iterative partial least squares method) was written for the analysis of spectra in Python in order to implement PCA, and the kNN algorithm (sklearn library) was also used.

The IR spectrum of each sample was combined into a single matrix, 70% of which data was then used to train the algorithm (training set), and 30% was used to test the algorithm's operability (test set). The training of the classification model over the entire range (4000-650) cm^{-1} did not lead to the desired result, so the combination of PCA and kNN was applied sequentially to each range with a length of 200 cm^{-1}, starting from 900 cm^{-1} and in increments of 200 cm^{-1}. The search was carried out for the maximum value of Macro Recall, that is, the average value of the proportion of correctly identified cases for each of the two classes. Macro Recall was evaluated by four applications of the algorithm on the same range using cross-validation.

As a result of the search, several suitable neighboring ranges were selected in the range from 3100 cm^{-1} to 3700 cm^{-1}, so the final training of the model took place at this interval. The training set included 14 samples of blood serum from donors and 5 samples of blood serum from patients with CLL. The test set included 6 donor samples and 3 CLL samples.

The algorithm correctly classified all analyzed samples of healthy donors and patients with CLL. Such a result, however, could have been caused by a successful choice of the training set, therefore, cross-validation was additionally carried out with the division of the original set into four equal parts and the Macro Recall averaged over four repetitions was evaluated. The Macro Recall value during cross-validation was 0.85, which means that there is an 85% probability that the new sample will be classified correctly.

Additionally, it was studied how the method works on the spectra of patients with MM. Six data sets were compiled - the second derivatives of the spectrum in the range of 1700-1350 cm-1 (set No. 1), spectra and their second derivatives in the range of 1700-1600 cm-1 (sets No. 2 and No. 3, respectively), the first and second main components, as well as their totality (sets No. 4, No. 5 and No.6). As a training sample, 20% of the total number of samples were used, which were randomly selected each time during 100 iterations of the classification. The accuracy of kNN averaged over 100 iterations for MM patients in

3 out of 6 sets approaches 100% (sets No. 3, No. 4 and No.6), and in 2 out of 6 it was 90% (sets No. 1 and No.2). That is, the combination of PCA and kNN methods also works well on the classification of MM patients.

The obtained result suggests that both donors and patients with CLL fall into their group with a probability of 85%. It is possible to increase the value of the Macro Recall metric, and hence improve the quality of classification, by increasing training and testing sets by obtaining new blood serum samples from healthy donors and patients with CLL. Good accuracy of classification of kNN and PCA on samples of MM patients was also shown. In the future, the considered algorithm can be used to create a CLL screening system, as well as MM screening system, which in turn is likely to increase the life span of patients with these diseases.

Acknowledgements. *Blood serum samples were kindly provided by the Federal State Institution "Russian Research Institute of Hematology and Transfusiology of the Federal Medical and Biological Agency" (St. Petersburg). Part of the work was carried out using the equipment of the St. Petersburg State University Scientific Park (Optical and Laser Methods of Substance Research, the Center for Diagnostics of Functional Materials for Medicine, Pharmacology and Nanoelectronics, Cryogenic Department).*

1. Mikhailets E.S. et al. Protein secondary structure analysis of serum from patients with oncohematological diseases // Journal of Physics Conference Series. 2021. V. 2103. No. 1. P. 012053.
2. Butyaev R. V. et al. Application of the principal component method for the analysis of IR spectra of blood serum of patients with oncohematological diseases // TOPICAL ISSUES of BIOLOGICAL PHYSICS AND CHEMISTRY, Volume 7, No. 3, 2022, pp. 462-466.
3. Chernyshev D. A. et al. Features of IR spectra of blood serum of patients with multiple myeloma. // Optics and Spectroscopy. 2023. Vol. 131, No. 6, pp. 805-809.
4. Telnaya E. A. et al. Infrared spectroscopy of blood serum of patients with oncohematological diseases // Biophysics. 2020. Vol. 65. No. 6. pp. 1154-1160.