PROCEEDINGS OF SCIENCE

# Taking into Account Mutual Correlations during Selection of Significant Input Features in Neural Network Solution of Inverse Problems of Spectroscopy

**N.O. Shchurov[a,b,1,*]; I.V. Isaev[a,c,2]; S.A. Burikov[a,b]; T.A. Dolenko[a,b]; K.A. Laptinskiy[a]; S.A. Dolenko[a,3]**

a. *D.V. Skobeltsyn Institute of Nuclear Physics, M.V. Lomonosov Moscow State University, 1/2 Leninskiye Gory, Moscow, 119991, Russia*

b. *Faculty of Physics, M.V. Lomonosov Moscow State University, 1/2 Leninskiye Gory, Moscow, 119991, Russia*

c. *Kotelnikov Institute of Radio Engineering and Electronics, Russian Academy of Sciences 11/7 Mokhovaya st., Moscow, 125009, Russia*

*E-mail:* [1] *Shchurov_no2@mail.ru,* [2] *isaev_igor@mail.ru,* [3] *dolenko@srd.sinp.msu.ru*

In the neural network solution of many physical problems, it becomes necessary to reduce the dimension of the input data in order to achieve a more accurate and stable solution while reducing computational complexity. When solving the inverse problem of spectroscopy, high multicollinearity between input features is often observed, as spectral lines may be much wider than the spectral channel width. This leads to the need to use a feature selection method that takes into account this characteristic. The method discussed in this article is based on iterative selection of input features with the highest Pearson correlation with the target variable and elimination of input features with high cross-correlation. This study compares the quality of the neural network solution to the problem of determining the concentration of heavy metal ions in water by Raman and absorption spectra on the full feature set and on its subsets produced by the considered feature selection method and by conventional methods of selection of significant input features.

**Keywords:** dimensionality reduction, spectroscopy, multicollinearity, feature selection

*The 6th International Workshop on Deep Learning in Computational Physics (DLCP2022)*
*6-8 July 2022*
*JINR, Dubna, Russia*

[*] Speaker





PoS(DLCP2022)026

## 1. Introduction

In the neural network solution of many physical problems, it becomes necessary to reduce the dimension of the input data [Chang, 2021]. This allows a more accurate and stable solution to be achieved while reducing computational complexity. Also, data processing increases the generalizing ability of the model.

When solving the inverse problem of spectroscopy, multicollinearity is often observed. Because of this, it is necessary to use a selection method that takes into account the correlation between input features. There are many approaches to detecting multicollinearity and various ways to solve this problem [Askin, 1982; Belsley, 2005]

Feature selection is an approach that requires selecting the most important subset of features for the target concept by removing redundant and irrelevant features. For feature selection in the case of high-dimensional data, filtering methods are often used [Chandrashekar, 2014]. The approach discussed in this article is to consider a special method of filter type. It is based on iterative selection of features with the highest Pearson correlation with the target variable and on elimination of features with high cross-correlation.

The paper compares the quality of the neural network solution to the problem of determining the concentration of heavy metal ions in water from Raman and absorption spectra on the full set of input features and on its subsets. These feature subsets are compiled using the selection method under consideration, as well as using traditional methods for selecting significant input features, such as cross-correlation based feature selection.

## 2. The objective of the study

The objective of the study is to test the effectiveness of the method of selecting essential features, taking into account multicollinearity in the case of solving the inverse problem of spectroscopy. In addition, the paper will consider the determination of the optimal input parameters of the algorithm and will compare the results obtained using this method with the results when selecting features based on cross-correlation and when training neural networks on a full data set.

## 3. Problem statement

As data for training the neural network, we used the experimentally obtained absorption and Raman spectra of solutions [Isaev, 2022]. The initial dataset contained 3806 patterns corresponding to solutions with different qualitative and quantitative composition. The considered solutions contained from 1 to 6 salts and from 2 to 5 ions in the concentration range of 0–0.14 M in increments of 0.01 M. The parameters determined in this problem were the concentrations of five ions and the pH value: $Cu^{2+}$, $Ni^{2+}$, $Co^{2+}$, $SO_4^{2-}$, $NO_3^{-}$, pH.

The intensity values of the spectrum in the corresponding channels were input features for the neural network. Since the bands of the spectra are wide, therefore the intensities of adjacent channels carry similar information. The input dimension of the problem was high and amounted to:

- 2048 features for Raman spectrum;
- 811 features for absorption spectrum.

## 4. The use of neural networks

To reduce the output dimension of the problem, the neural network solution used autonomous parameterization, that is, training occurred for six separate single-output networks – one for each determined parameter.

- To prevent overfitting, the early stopping method was used, so the initial dataset was divided into subsets in the following ratio:
- Training 70 % 2 656 patterns
- Validation 20 % 750 patterns
- Test 10 % 400 patterns
- The neural networks used had the following training parameters:
- Architecture: multilayer perceptron with one hidden layer
- Number of neurons in the hidden layer: 32
- Activation function:
  - Hidden layer - sigmoid
  - Output layer - linear
- Early stopping: after 800 epochs with no improvement on the validation set

To reduce the impact of the initial weight initialization on the results, we trained each neural network 5 times with different initial weights and then averaged the results.

## 5. Description of the iterative feature selection algorithm

This article discusses an iterative method based on feature selection (IFS) and taking into account multicollinearity. Further in this article, correlation means the Pearson correlation.

First, the algorithm selects the feature with the highest correlation with the target variable. Then, all features whose correlation with the one chosen at the previous stage was higher than some threshold value are excluded from the set.

This process is repeated either until there are no features in the initial set whose correlation value with the target variable is greater than a certain threshold, or until the features run out (Fig. 1).

Thus, this method has two parameters that have to be set:

1. The minimum allowable value of correlation with the target variable (hereinafter referred to as the XY threshold)

2. The maximum allowable value of correlation with other input features (hereinafter referred to as the XX threshold)

A similar feature selection method is discussed in the article [Biesiada1, 2007].

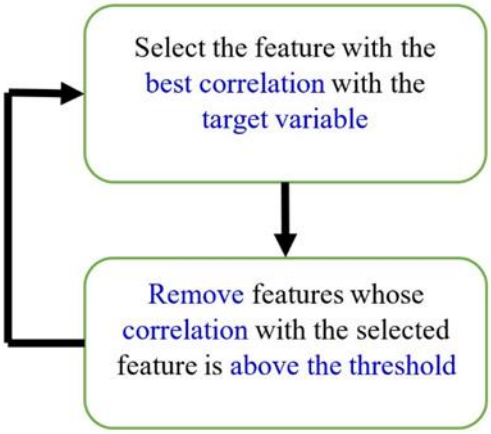


*Fig. 1. Scheme of the iterative feature selection algorithm*

## 6. Application of the iterative feature selection algorithm

For the spectroscopy data of the solutions, the absolute value of Pearson correlation between each pair of the input features was calculated, the results are presented in the form of heat maps (Fig. 2).

It can be seen from the heat maps that a large number of input features, especially those located in neighboring channels of the spectrum, indeed have high correlation values with each other. This may indicate the redundancy of some features, which we will exclude from the data set using the feature selection method.

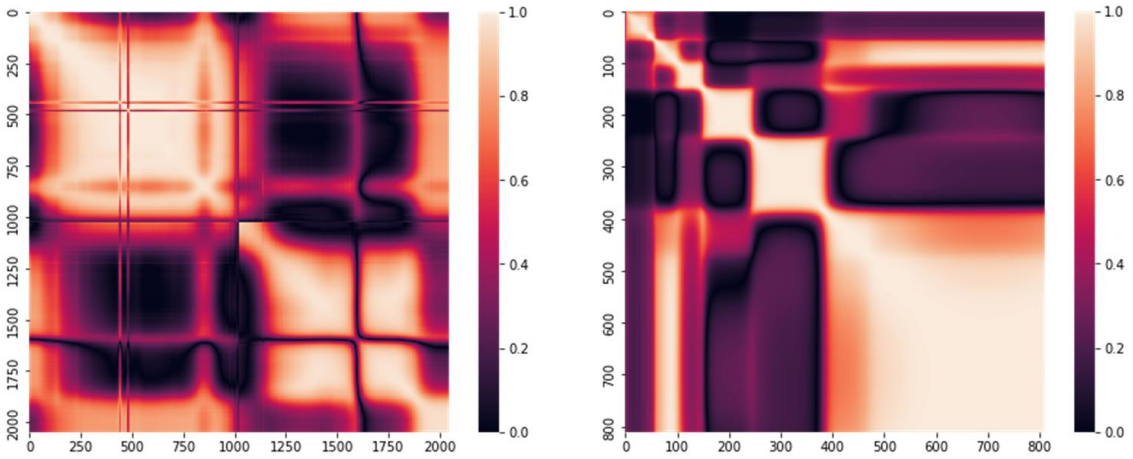

*Fig. 2. The heat maps of absolute values of correlation between input features for Raman spectroscopy (left) and for absorption spectroscopy (right).*

As it was mentioned above, the method discussed in this article has two threshold values to be set. We chose the XY threshold to be 0.3. For the correct choice of the threshold XX, graphs of the dependence of the number of selected features on the value of the threshold XX were plotted (Fig. 3). For further consideration, we took four threshold values from the inflection segment of the graph: XX = 0.80, 0.85, 0.90 and 0.95.

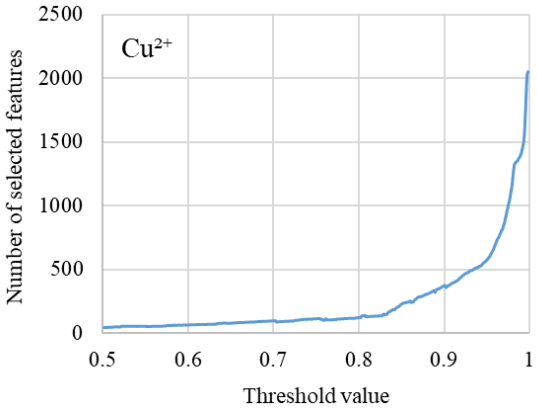


*Fig. 3. Dependence of the number of selected features on the XX-threshold value.*

An example of features selected by this algorithm is shown as black dots on the plot of the Raman spectrum and the absorption spectrum of three different solutions displayed on the graph in different colors (Fig. 4). It can be seen from the graph that for the Raman spectrum, the algorithm selected features corresponding to the characteristic bands of ions (channels 450-500) and stretching vibrations of water (channels 1500-1700). This choice corresponds to the physical concept of the dependence of the shape of the spectrum on the composition of the solution. For the absorption spectrum, the algorithm selected features somewhat shifted from the maxima of the spectral bands. Such a choice can be explained by the fact that the bands overlap each other, and with such a choice of features, the influence of various elements is most distinguishable.

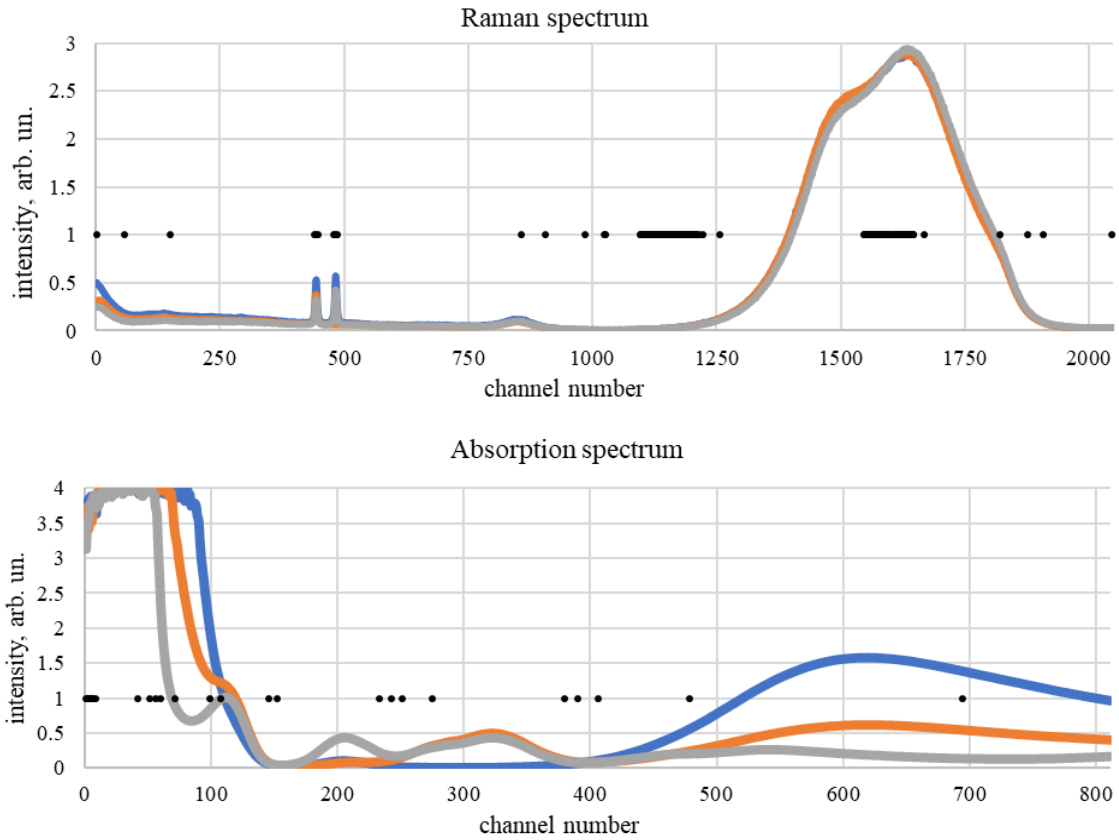


*Fig. 4. An example of features selected by the iterative algorithm (black dots) for Raman spectroscopy (top) and for absorption spectroscopy (bottom).*

## 7. Results of solving the inverse problem of spectroscopy

To train the neural networks, features were selected using various threshold values (XX = 0.80, 0.85, 0.90, 0.95). The quality of neural networks increases with the XX threshold value, which can be caused by reaching the optimal value for the considered dataset, as well as by the increase in the total number of selected input features. Based on the quality of the obtained models (Fig. 5), we chose the threshold value XX = 0.95 for further use.

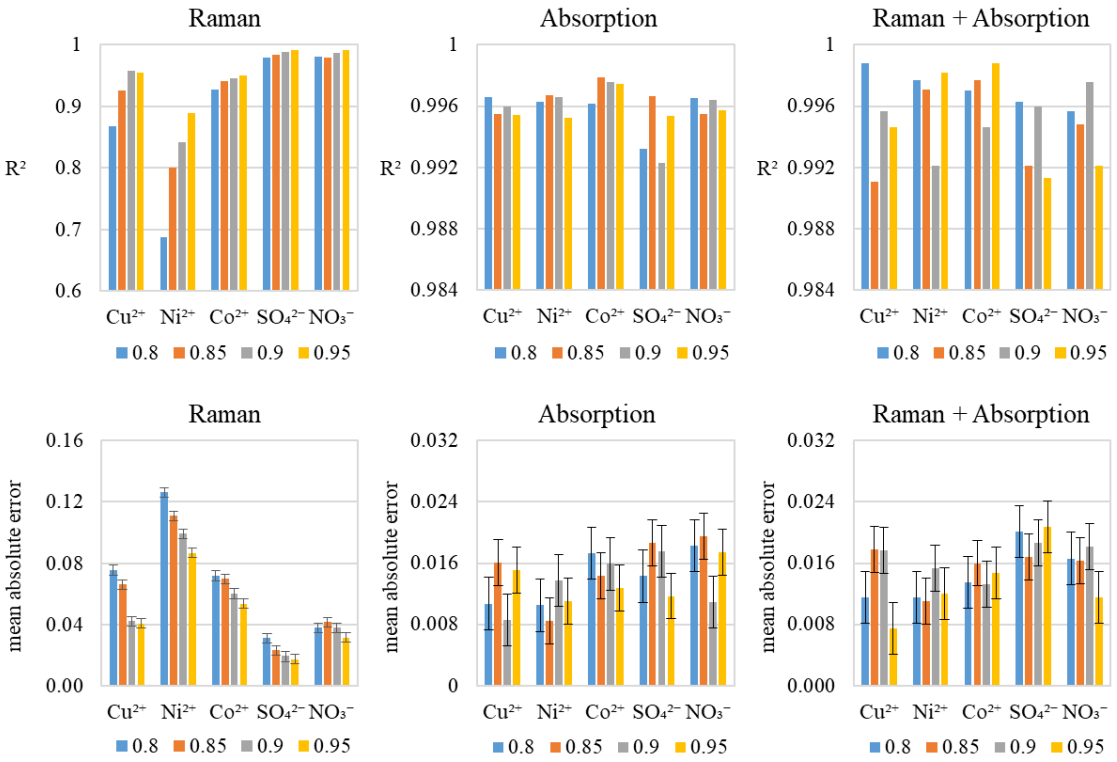


*Fig. 5. The quality of the solution (top – R², bottom – mean absolute error) of the inverse problem for various XX-threshold values in the iterative feature selection algorithm and for various input data.*

The common feature selection method based on cross-correlation is inferior in the quality of solution to the iterative feature selection algorithm, and sometimes it does not achieve an acceptable result at all (Fig. 6). This can be explained by the fact that such an algorithm uses a large number of features from neighboring channels of the spectrum, since the multicollinearity is high. As a result, the selected set consists of channels belonging to only one or several spectral lines.

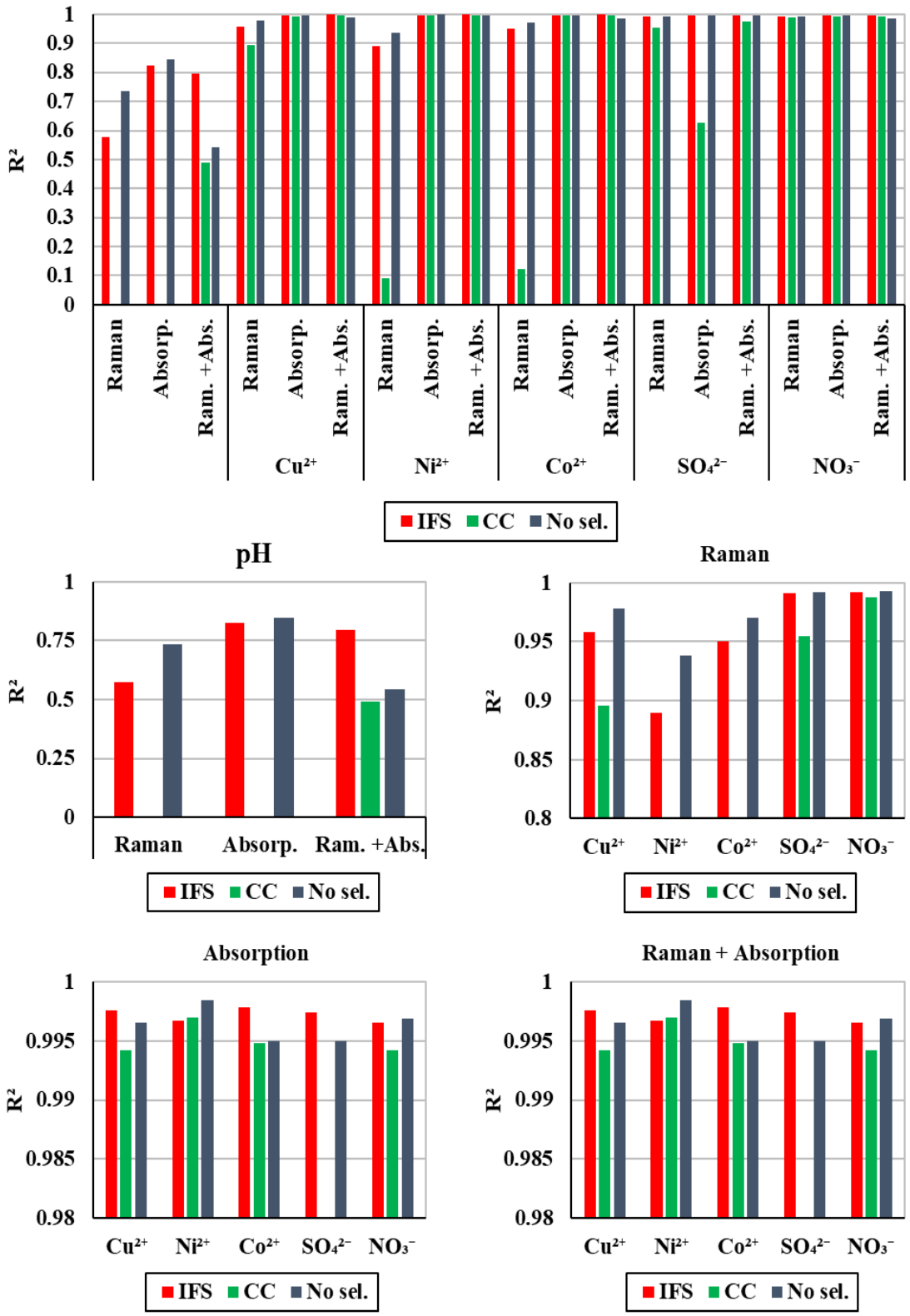


*Fig. 6. The quality of the solution ($R^2$) of the inverse problem for the iterative feature selection algorithm, cross-correlation feature selection and on the full set of input features for various input data. Top – summarized results, bottom – detailed results.*

When training neural networks on the full data set, combining spectra leads to a decrease in the quality of the solution (Fig. 6). This is due to too many irrelevant features [John, 1994]. In [Isaev, 2022], the absorption and Raman spectra were also joined, and neural networks showed poorer results on the joint set compared to using only the absorption spectrum.

The method considered in the study, on the contrary, makes it possible to improve the quality of the solution on the combined set, and the value of $R^2$ in this case is higher than when training the neural network on the full data set (Fig. 6).

The greatest advantage when using the iterated feature selection method appears when the Raman and absorption spectra are combined into one set (Fig. 6). This can be explained by the fact that connecting an additional data set in the case of using this method does not lead to a large increase in the dimension of the problem.

When determining the pH of a solution, the iterated feature selection method shows an advantage on the combined dataset, but when using only the Raman spectrum or absorption, the quality of networks trained on the full dataset is higher (Fig. 6).

## 8. Conclusions

The features selected using the considered method correspond to physical representations and allow one to reduce the dimension of the input data by an order of magnitude. There is also an improvement in the quality of the neural network solution when combining Raman and absorption spectroscopy data, which makes it possible to use this method to select the most significant features from data sets of various kinds.

## Acknowledgement

This study has been performed at the expense of the Russian Science Foundation, grant no. 19-11-00333, https://rscf.ru/en/project/19-11-00333/.

## References

*Askin R. G.* Multicollinearity in regression: Review and examples // Journal of Forecasting. — 1982. — Vol. 1. — №. 3. — P. 281-292.

*Belsley D. A.,* Kuh E., Welsch R. E. Regression diagnostics: Identifying influential data and sources of collinearity. — John Wiley & Sons, 2005.

*Biesiada J.,* Duch W. Feature selection for high-dimensional data — a Pearson redundancy based filter // Computer recognition systems 2. — Springer, Berlin, Heidelberg, 2007. — P. 242-249.

*Chandrashekar G.,* Sahin F. A survey on feature selection methods // Computers & Electrical Engineering. — 2014. — Vol. 40. — №. 1. — P. 16-28.

*Chang L.*, Wang J., Woodgate W. Analysing spectoscopy data using two-step group penalized partial least squares regression // Environmental and Ecological Statistics. — 2021. — Vol. 28. — №. 2. — P. 445-467.

*Isaev I.,* Gadzhiev, I., Sarmanova, O., Burikov, S., Dolenko, T., Laptinskiy, K., Dolenko, S. Using method integration transfer learning for neural network solution of an inverse problem in optical spectroscopy // Laser Physics, Photonic Technologies, and Molecular Modeling. — SPIE, 2022. — Vol. 12193. — P. 217-222.

*John G. H.*, Kohavi R., Pfleger K. Irrelevant features and the subset selection problem // Machine learning proceedings 1994. — Morgan Kaufmann, 1994. — P. 121-129.