ADVANCES IN CYBER-PHYSICAL SYSTEMS

Vol. 9, No. 2, 2024

PREDICTING THE DURATION OF TREATMENT USING PERSONALIZED MEDICAL DATA

Mykola Stakhiv

Lviv Polytechnic National University, 12, Bandera Str, Lviv, 79013, Ukraine.
Author's e-mail*: mykola.a.stakhiv@lpnu.ua*

https://doi.org/10.23939/acps2024.02.146

Submitted on 25.09.2024

***Abstract*: The article describes the problem of data personalization by identifying the individual characteristics necessary to solve the personalization problem. The essence of the researched problem of personalization and the solution of tasks of the estimated correlation between individual characteristics and the solution using the forecasting model has been also highlighted. This study focuses on solving the problem of formalization of the studied object and the formalization of its conditions during treatment or rehabilitation, which will optimize the processes of treatment, analysis of individual patient characteristics, and forecasting possible personalized solutions for health care, focusing on patient health.**

Index Terms: information system, personalized medicine, modeling, ensemble of models.

I. INTRODUCTION

In recent years, paradigms and visions of medicine and healthcare have been changed, and the healthcare policy of Western countries is focused on disease prevention and awareness, which is confirmed by appropriate medical literacy. In addition, the European Commission develops and implements various programs to promote the widespread adoption of comprehensive digital solutions that will improve the quality of life of citizens [1], while demonstrating significant improvements in the efficiency of health services and care in Europe. Particularly important for the implementation of these programs are, among other measures, increasing the level of literacy of the population in the field of health care, the collection of medical, inpatient, and outpatient observations, and of fundamental importance – personalization in medicine [2]. Determining the necessary individual characteristics to solve the personalization problem depends on the key factors of object identification. For a formalized presentation of the researched object in medicine, the main parameters of its general condition with its defined characteristics are taken into account. The data of the research object, coming from various sources, are used for treatment for medical and health purposes. Nowadays, however, evidence supporting the provision of specialized medical care is largely concentrated and may be collected as part of disease screening or diagnostic procedures.

II. LITERATURE REVIEW AND PROBLEM STATEMENT

Personalized medicine, in the digital age, must be supported by small and big data, and artificial intelligence technology must be used to support risk prevention, prediction and detection, and medical intervention. In the work [3] the author emphasizes that the primary goal of personalized medicine is to enhance diagnosis and treatment for individual patients, aiming to integrate this approach across various medical specialties. Additionally, he highlights the importance of gathering diagnostic and treatment data for use in medical and bioinformatics processes, focusing on the role of big data in patient-specific insights. It is a well-known fact that artificial intelligence gives us many opportunities to understand and solve many complex problems in the practical and scientific space. Artificial intelligence can be used in systems designed to detect, track, and predict disease outbreaks. The more effectively the virus's spread is tracked, the faster and more efficiently it can be combated [4].

Data mining, including in medicine, is used by several scientists. For example, the work [5] demonstrates the actual effects of missing data for regression by analyzing its impact in several publicly available databases implementing popular algorithms like Decision Tree, Random Forests, Adaboost, K-Nearest Neighbors, Support Vector Machines, and Neural Networks. Similar results were obtained by Sokolova O. [6] and Sina Khanmohammadi [7] for associative classification of medical data and Anthony Costa Constantine for comprehensive survey [8] and data interviews for intelligent Bayesian models for medical decision support. The Bayesian network is also used to reformulate the Quick Medical Reference (QMR) model in decision theory. However, with the advent of big data technology, Bayesian networks have been slow. Therefore, the parallelization of Bayesian networks was developed in the work of Y. Tang [9]. Bayesian networks are also used to diagnose dementia, Alzheimer's disease, and mild cognitive impairment. The Bayesian belief network is also used in [10] and [11] for the analysis of aging diseases. However, even under conditions of parallelism, for multi-parameter, large-volume, and dynamic medical data, Bayesian networks should only be used in conjunction with other machine learning methods. The apparatus of artificial neural networks, including those using fuzzy logic, is also actively proposed for the analysis of various medical data. Thus, in the work of E. Bodyanskyi and I. Perova [12], a system of rapid medical diagnostics based on auto-associative neuro-fuzzy memory is proposed. This system is

characterized by the simplicity of the architectural solution and its software implementation and provides diagnosis of patients with several parameters.

III. SCOPE OF WORK AND OBJECTIVES

This study focuses on solving the problem of formalization of the researched object and formalization of its conditions during treatment or rehabilitation, which will allow for optimization of the processes of treatment, analysis of individual characteristics of patients, and prediction of possible personalized decisions regarding the provision of medical care, focusing on the assessment of the patient's health.

IV. PROBLEM STATEMENT

Next, the essence of the researched problem of personalization and the solution of tasks of the estimated correlation between individual characteristics and the solution using the forecasting model is highlighted. The experimental setup is organized as follows: exploratory data analysis (normalization and coding of features), conditional development of space, weak choice of predictors, hierarchical predictor of development, and evaluation of results.

All calculations are made using RStudio. Data were run through a Data Sampler for balancing. As a result, all cases are taken into account. This means that the collected data set is balanced.

Personalized patient data is processed from the collected data set. This data set was collected in the surgical department of the Lviv Public Hospital (Ukraine). The patients underwent clinical treatment for postoperative complications in the abdominal cavity.

The dataset consists of several attributes. Age is represented as an integer and serves as a time-dependent parameter and efficiency indicator (A_in). Gender is a logical value and acts as a time-independent parameter. Weight is a categorical variable that functions as a time-dependent parameter and performance indicator (A_in). The date of reception is recorded as a date attribute. Diagnostics, which is a categorical variable, is used as a time-independent parameter to select a protocol (PFSoi). The associated diagnosis is a time-dependent categorical parameter (RPFSoi), while flora is also a time-dependent categorical parameter (RPFSoi). The drug attribute is categorical and depends on the protocol (PFSoi), as does the active substance. Finally, time in the hospital is an integer that represents a time-dependent parameter (hospital bed days) and serves as the target parameter.

Each instance represents a GSo object. The task is to predict the number of days in the hospital (duration of treatment) based on medical treatment and personal parameters of the patient. The dataset consists of 51 instances and 10 parameters. *Time_in_hospital* is the target variable. After a preprocessing step and using one hot coding for the categorical variables, the dataset consists of 39 features. Missing values can be found in the Flora function and the Related Diagnostics function. The missing data imputation procedure is not used because the nature of the missing data is truly an empty value. The correlation between the parameters is shown in Fig. 1.

The following steps are taken to arrange the conditional space: selection of the most important features; and instances divided into clusters with similar time-dependent and time-independent parameters. The initial feature selection is carried out using the correlation matrix, Boruta, and regression tree. Hard voting is used to finalize the feature.

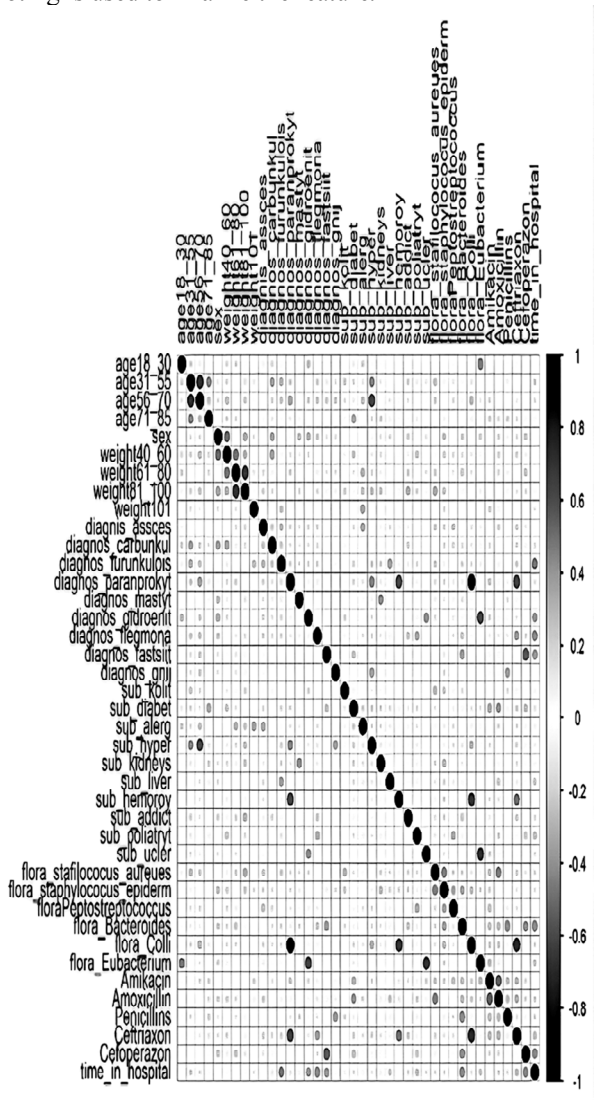

Fig. 1. Correlation matrix

No significant correlation is found. The following Boruta algorithm is used [13] as it is depicted in Fig. 2. The Boruta algorithm is a wrapper built based on the Random Forest Classification algorithm. Fig. 2 shows important (green), and indicative (yellow) boxes. The blue box plots correspond to the minimum, mean, and maximum Z score of the shadow attribute.

The important variables include the solution meanImp, along with the following confirmed diagnoses: diagnoses_furunkulois with a value of 18.631571, diagnose_gidroenit at 9.901603, diagnose_flegmona at 8.915723, and sub_diabetes at 7.557117.

The following regression tree is used for feature selection. It begins with a root node that contains 51 instances, with a predicted value of 1981.92200 and a

variance of 10.372550. The first branch evaluates the diagnosis of furunculosis, where instances with a value of 0.5 or greater include 11 instances, yielding a predicted value of 17.63636 and a variance of 4.81818. In contrast, instances with a value of less than 0.5 consist of 40 instances with a predicted value of 1531.6000 and a variance of 11.90000.

Within this branch, if sub-diabetes is less than 0.5, there are 33 instances with a predicted value of 1075.87900 and a variance of 11.060610. Further, if ceftriaxone is less than 0.5, this group contains 26 instances, resulting in a predicted value of 818.65380 and a variance of 10.115380. Among these, when gender is 0.5 or greater, there are 12 instances with a predicted value of 602.66670 and a variance of 8.666667. Conversely, if gender is less than 0.5, there are 14 instances with a predicted value of 169.21430 and a variance of 11.357140.

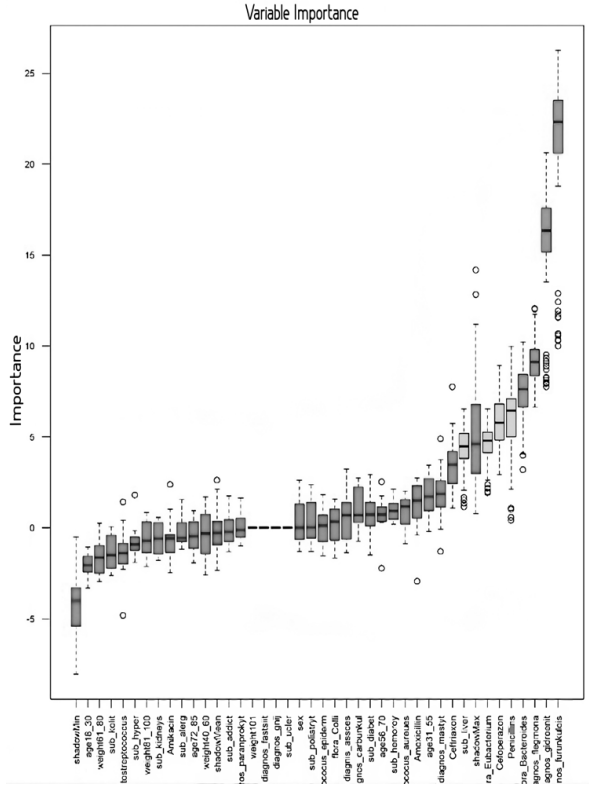

Fig. 2. The result of the Boruta algorithm

If ceftriaxone is 0.5 or greater, there are 7 instances with a predicted value of 147.71430 and a variance of 14.571430. Lastly, for instances where sub-diabetes is 0.5 or greater, there are 7 instances with a predicted value of 322.85710 and a variance of 15.857140.

The above regression tree structure allows for finding significant attributes and the values of those attributes. The regression tree is built on the following features: diagnosis of furunculosis, sub-diabetes, ceftriaxone, and sex.

The key variables are similar for both methods. The hard vote for the three feature selectors includes the following: diagnosis of furunculosis, sub-diabetes, diagnosis of hidradenitis, ceftriaxone, diagnosis of phlegm, and gender.

Ceftriaxone treatment affects the duration of hospital stay. Therefore, only time-dependent parameters are important for predicting hospital stays.

Next, clustering is used. Visual assessment of (cluster) tendency (VAT) is used to analyse the possibility of splitting objects. VAT shows a poor tendency towards clustering (Fig. 3). Small differences are represented by dark shades, and large ones by light ones [14].

The same result is observed with the k-means visualization (Fig. 4). There is an overlap of clusters, especially for clusters one and three.

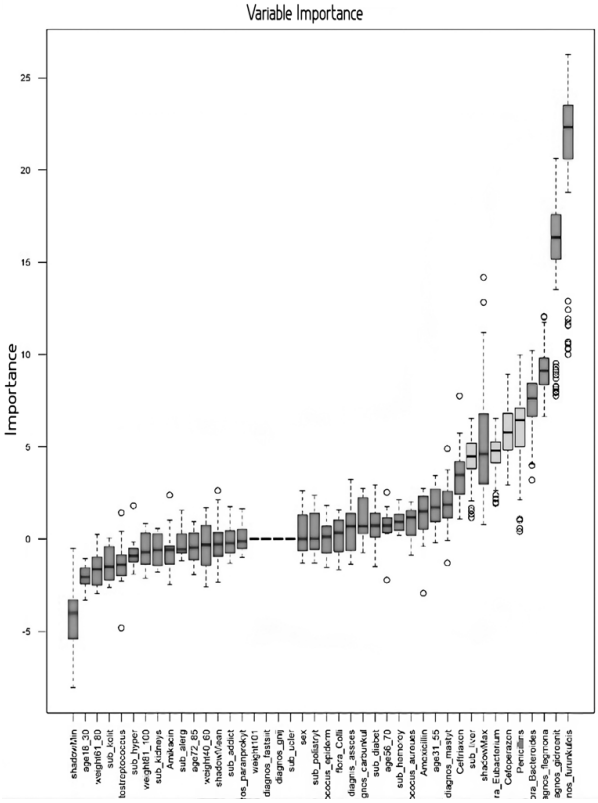

Fig. 3. VAT usage results

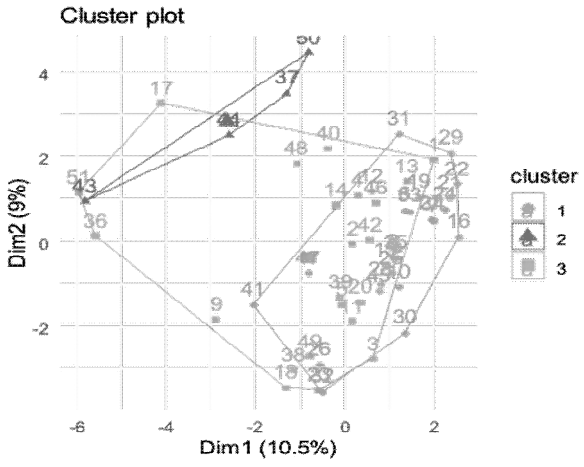

Fig. 4. Clustering results using k-means

The Hopkins statistic [15] is 0.71. This means that the data set does not cluster very well. Objects 7, 13, 17, 20, 26, 34, 39, 42, 45 should be analysed separately. The value of the membership function for these objects is less than 0.65. That is why strong differences between objects are impossible to find.

In the next step, several predictors are built. Two metrics are considered: root mean square error (RMSE) and mean absolute percentage error (MAPE). The results are shown in Table 1.

In the next step, weak predictors will be used for each cluster. Instances with an unknown cluster were added to the fourth cluster. A total of 4 clusters are taken into account.

A hierarchical predictor is built by first using fuzzy k-means to divide the objects into four clusters, as it is shown in Figure 4. For each cluster, random forest, linear regression, SVM with a radial basis kernel, and SVM with a polynomial kernel are applied individually. The final prediction is determined by calculating an average vote from the results obtained, with the average value being selected.

The prediction accuracy of the hierarchical predictor is presented in Table 3. The quality is worse than for the entire data set.

Table 1

Results of predictors

Model	RMSE	MAPE
Linear regression	6.1097	0.4946
Regression tree	4.9706	0.4763
Random forest (500 trees, mtree-3)	3.5604	0.3753
knn	3.3604	0.3753
SVM with Radial Basis kernel	3.1946	0.2923
SVM with Polynomial kernel	2.2621	0.2670
ANN with 12 units in single hidden layer	2.0972	0.2025

Table 2

Prediction accuracy for selected traits

Model	RMSE	MAPE
Linear regression	3.9725	0.3240
Regression tree	4.9706	0.4763
Random forest (500 trees, mtree-3)	3.5464	0.2730
knn	3.3464	0.2730
SVM with Radial Basis kernel	3.0952	0.2386
SVM with Polynomial kernel	2.2269	0.1766
ANN with 12 units in single hidden layer	2.0598	0.1513

Table 3

The predictive accuracy of the proposed predictors

Model based on whole variables	RMSE	MAPE
Hierarchical predictor	1.4012	0.1377

The analysis with the selected variables is given below in Table 2.

K-fold repeated cross-validation is used to validate small data sets. The advantage of this technique is the possibility of parallelization. The result is presented in Table 4.

Table 4

Re-cross-validation of K-crats

Model based on selected variables	RMSE	MAPE
Hierarchical predictor	1.4012	0.1029
Hierarchical predictor with repeated K-fold cross-validation, 5-fold, repeated 3 times	1.4011	0.1027

It can be observed that the RMSE for the hierarchical predictor database on the entire dataset is only slightly higher than for the selected features. On the other hand, the prediction accuracy for split clusters is better than the best weak predictor Artificial Neural Network (ANN) with 12 units in one hidden layer. The accuracy of the hierarchical predictor with multiple cross-validation of K-folds is not that much higher than that of the non-cross-validated method.

The obtained results of the study of the proposed models for the collected, in general, confirm the hypothesis of differences in the accuracy of prediction for the entire data set and the conditional space built on clustering and feature selection. However, two important remarks should be made immediately. Firstly, the number of instances in the collected dataset is too small, and this hypothesis needs to be validated using a larger dataset. Efforts are currently underway to expand the dataset. Nonetheless, a similar approach was successfully applied in a previous study on a different dataset [15-16], yielding the same imputation results. Secondly, the authors suggest that the prediction accuracy is likely to be significantly influenced by the number of missing values.

The difference between the best weak predictor (perceptron) is 1.01 better for individual features (Table 1 and Table 2). The quality of the developed hierarchical predictor for the RMSE metric is 1.47 better than that of the best weak predictor. The regression tree shows the same result for the entire data set and the selected variables. Linear regression shows a 1.53 better RMSE for selected objects compared to the entire data set. The remaining weak predictors show better results on the selected features.

The choice of function is important not only for increasing accuracy. The training time of the hierarchical predictor for the selected features is 1.2 times less than for the entire data set.

The proposed method is used for small data sets and is similar to [17]. Here, the authors propose to improve the RBF-based input doubling method by introducing additional elements into the formula for calculating the output signal of the method. The accuracy is improved by 4% in terms of MAE and RMSE compared to the basic method.

V. CONCLUSION

This article presented the selection of features and bed days in hospital forecasting using conditional space and developed hierarchical predictors. A dataset of personalized medical parameters was collected at a public hospital. This

data set was used to predict the number of hospital days. The patients underwent clinical treatment for postoperative complications in the abdominal cavity. The study showed a low pairwise correlation between a huge subset of the parameters listed in the data set. However, to improve the quality of the predicted model, a proper selection of the function was required.

Data preprocessing improved the quality of the analysis. Boruta, regression tree, and correlation were used for feature selection. Selection results were formed based on strict voting of all feature selectors. Several clustering algorithms was used to divide the object into separate clusters. This splitting allowed for the development of a conditioned space based on time-dependent data and medical protocol.

In this work, a hierarchical predictor was developed based on a combination of clustering results and four weak predictors for each cluster separately. Thus, the proposed algorithm showed a higher prediction accuracy than the best predictor perceptron.

References

[1] Tresa, E., Czabanowska, K., Clemens, T., Brand, H., Babich, S. M., Bjegovic-Mikanovic, V., & Burazeri, G. (2022). Europeanization of health policy in post-communist European societies: Comparison of six Western Balkan countries. *Health policy (Amsterdam, Netherlands), 126*(8), 816–823. DOI: *https://doi.org/10.1016/j.healthpol.2022.05.01*.

[2] Malek N. P. (2017). Personalisierung in der Medizin der Zukunft: Chancen und Risiken [Personalization in the medicine of the future : Opportunities and risks]. *Der Internist, 58*(7), 650–656. DOI: *https://doi.org/10.1007/s00108-017-0265-5*.

[3] Djulbegovic, B., & Guyatt, G. H. (2017). Progress in evidence-based medicine: a quarter century on. *The Lancet*, *390*(10092), 415-423. DOI: *https://doi.org/10.1016/s0140-6736(16)31592-6*.

[4] Danhof, M., Klein, K., Stolk, P., Aitken, M., & Leufkens, H. (2018). The future of drug development: the paradigm shift towards systems therapeutics. *Drug discovery today*, *23*(12), 1990-1995. DOI: *https://doi.org/10.1016/j.drudis.2018.09.002*.

[5] Marcelino, C. G., Leite, G. M. C., Celes, P., & Pedreira, C. E. (2022). Missing Data Analysis in Regression. *Applied Artificial Intelligence, 36*(1). DOI: *https://doi.org/10.1080/08839514.2022.2032925*.

[6] Mishyna, M., Volokh, O., Danilova, Y., Gerasimova, N., Pechnikova, E., & Sokolova, O. S. (2017). Effects of radiation damage in studies of protein-DNA complexes by cryo-EM. *Micron (Oxford, England: 1993), 96*, 57–64. DOI: *https://doi.org/10.1016/j.micron.2017.02.004*.

Mykola Stakhiv, was born in Lviv, Ukraine, in 1996. Starting from 2023 he became a Postgraduate student of the Department of System Analysis of Lviv Polytechnic National University, Lviv, Ukraine, Computer vision. His research interests include methods of analyzing text flow, big data mining, unsupervised machine learning, intelligent systems, Support Decision Making Systems.

[7] Khanmohammadi S. (2017). An improved synchronization likelihood method for quantifying neuronal synchrony. *Computers in biology and medicine, 91*, 80–95. DOI: *https://doi.org/10.1016/j.compbiomed.2017.09.022*.

[8] Perov, Y.N., Graham, L., Gourgoulias, K., Richens, J.G., Lee, C.M., Baker, A., & Johri, S. (2019). MultiVerse: Causal Reasoning using Importance Sampling in Probabilistic Programming. *Symposium on Advances in Approximate Bayesian Inference*. DOI: *https://doi.org/10.48550/arXiv.1910.08091*.

[9] Tang, Y., Wang, J., Nguyen, M., & Altintas, I. (2019). PEnBayes: A Multi-Layered Ensemble Approach for Learning Bayesian Network Structure from Big Data. *Sensors (Basel, Switzerland), 19*(20), 4400. DOI: *https://doi.org/10.3390/s19204400*.

[10] Lakho, Shamshad & Jalbani, Dr & Vighio, Muhammad & Memon, Imran & Siraj, Saima & Soomro, Qamar Un Nisa. (2017). Decision Support System for Hepatitis Disease Diagnosis using Bayesian Network. *Sukkur IBA Journal of Computing and Mathematical Sciences*. DOI: *https://doi.org/https://doi.org/10.30537/sjcms.v1i2.51*.

[11] Kallen, V., Tahir, M., Bedard, A., Bongers, B., van Riel, N., & van Meeteren, N. (2021). Aging and Allostasis: Using Bayesian Network Analytics to Explore and Evaluate Allostatic Markers in the Context of Aging. *Diagnostics (Basel, Switzerland), 11*(2), 157. DOI: *https://doi.org/10.3390/diagnostics11020157*.

[12] Perova, Iryna & Bodyanskiy, Yevgeniy. (2017). Fast medical diagnostics using autoassociative neuro-fuzzy memory. *International Journal of Computing, 16*, 34-40. DOI: *https://doi.org/10.47839/ijc.16.1.869*.

[13] Anand, Neeyati & Sehgal, Riya & Anand, Sanchit & Kaushik, Ajay. (2021). Feature selection on educational data using Boruta algorithm. *International Journal of Computational Intelligence Studies. 10*(1), 27-35. DOI: *https://doi.org/10.2478/logi-2024-0008*.

[14] Wang, M., Abrams, Z. B., Kornblau, S. M., & Coombes, K. R. (2018). Thresher: determining the number of clusters while removing outliers. *BMC bioinformatics, 19*(1), 9. DOI: *https://doi.org/10.1186/s12859-017-1998-9*.

[15] Melnykova, Nataliia & Shakhovska, Natalya & Greguš, Michal & Melnykov, Volodymyr. (2019). Using Big Data for Formalization the Patient's Personalized Data. *Procedia Computer Science, 155*, 624-629. DOI: *https://doi.org/10.1016/j.procs.2019.08.088*.

[16]Shakhovska, N., Izonin, I., & Melnykova, N. (2021). The Hierarchical Classifier for COVID-19 Resistance Evaluation. *Data, 6*, 6. DOI: *https://doi.org/10.3390/data6010006*.

[17] Izonin, I., Tkachenko, R., Dronyuk, I., Tkachenko, P., Gregus, M., & Rashkevych, M. (2021). Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. Mathematical biosciences and engineering: *MBE, 18*(3), 2599–2613. DOI: *https://doi.org/10.3934/mbe.2021132*.