Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 4268–4275

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

COVID-19 antibody level analysis with feature selection approach

Wiesław Paja[a,*], Krzysztof Pancerz[b], Catalin Stoean[c]

[a] College of Natural Sciences, University of Rzeszów, Rzeszów, Poland
[b] Institute of Technology and Computer Science, Academy of Zamosc, Poland
[c] University of Craiova, Department of Computer Science, Craiova, Romania

Abstract

The study presented here considers the analysis of a medical dataset for the identification of the stage of onset of COVID-19 coronavirus. These data, presented in previous work by the authors, have been subjected to extensive analysis and additional calculations. The data were obtained by analyzing blood samples of infected individuals at 1, 3, and 6 months after COVID-19 infection. Results were obtained from FTIR spectrometry experiments. The results indicate a very effective ability to identify the different states of infection, and between 1 and 6 months even perfect. Specific spectrometry wavelength ranges can also be distinguished as medical markers.

Keywords: COVID-19; FTIR; Fourier Transform Infrared spectrometry; feature selection; computer aided medical diagnosis;

1. Introduction

The development of new computer-based methods to aid the diagnosis of various types of medical conditions has been an important stream of research in recent years. Such methods have been applied in many medical issues [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. The development of information technologies, especially machine learning and artificial intelligence ones, causes the emergence of new areas in which they can be applied. Such methods are also combined with other physical, chemical or biotechnological approaches. One challenge that has recently appeared is to develop effective methods to diagnose and treat COVID-19 coronavirus.

Coronavirus disease, COVID-2019, is an acute respiratory infectious disease caused by SARS-CoV-2 virus infection firstly recognized in November 2019, in central China (Wuhan city, Hubei province) during a series of illnesses that initiated a pandemic of the disease [13]. The standard method for diagnosing infection is a real-time reverse transcriptase polymerase chain reaction (RT-qPCR, real-time RT-PCR) test performed from a nasopharyngeal swab

* Corresponding author. Tel.: +48-502-449-825.
E-mail address: wpaja@ur.edu.pl

1877-0509
10.1016/j.procs.2022.09.490

or sputum sample, which provides results in a few hours to two days. Antibody analysis from a blood serum sample can also be used as a diagnostic method, providing a result within a few days. The disease can also be diagnosed by evaluating a combination of symptoms, risk factors, and the result of a chest CT scan showing features of pneumonia [14].

Fourier Transform Infrared (FTIR) spectroscopy is an analytical methodology used in industrial and academic laboratories to study the structure of individual molecules and the composition of molecular mixtures. FTIR spectroscopy uses modulated mid-infrared energy to examine samples. Infrared light is absorbed at specific frequencies directly related to the vibrational energies of the bonds between atoms in a molecule. When the vibrational bond energy and the mid-infrared light energy are equal, the bond can absorb this energy. Different bonds in a molecule vibrate with different energies and therefore absorb different wavelengths of infrared radiation. The position (frequency) and intensity of these individual absorption bands make up the overall spectrum, creating a characteristic ”fingerprint” of the molecule.

Previous works [1, 2] have confirmed that the Fourier-transform infrared spectroscopy (FTIR) and the Raman spectroscopy can be used to detect COVID-19. In addition to the ability to distinguish the blood serum of individuals with COVID-19 from that of healthy individuals, the main focus has been on the use of spectroscopy methods to predict the timing of SARS-CoV-2 antibody emergence, to estimate the risk of reinfection and the need for vaccination by periodically assessing total antibody levels in individuals who have recovered from COVID-19. This article presents an extended analysis of the wave absorption data used, with particular emphasis on the selection of relevant wavebands responsible for identifying diseased cases.

2. Materials and methods

The dataset contains 47 cases of patients ranging in age from 26 to 77 years. These are health care workers infected with COVID virus. Infections were confirmed by SARS-CoV-2 RT-PCR test. Additional patient data, such as week, sex, symptoms, so-called comorbidities, are described in detail in [1]. Blood samples from each patient were collected at specific intervals: at 1 month after infection (M1), at 3 months (M3) and the last one at 6 months (M6). The number of patients is 47, each patient was studied at 1, 3 and 6 months of illness. Hence, the dataset contained approximately 141 learning cases each described by 156 or 166 attributes (wavenumbers) depending on the range. To distinguish between patient groups, 2 two-class subsets were created containing individuals at 1 and 3 months (Group M1 and M3) and 1 and 6 months of illness (Group M1 and M6). Each contained approximately 81-85 cases. In machine learning, we use the Leave-One-Out Cross Validation (LOOCV) approach in such a case. When the number of cases in the dataset is less than 100, we perform as many model building processes as there are learning cases, and there is one test case in each process. In the process of building learning models, six machine learning algorithms with different operating methodologies were used: C5.0 Decision Trees [15], Random Forest [16], Deep Neural Networks [17], k Nearest Neighbor clustering [18], XGBoost trees [19], and Support Vector Machine [20]. The results of the analysis of these subsets are included in Tables 1 and 2. In addition, a feature (wavenumbers) selection process [8, 21] of relevance was performed and thus the number of wavenumbers was limited to only relevant ones. The results for these subsets are also included in Tables 1 and 2.

3. Results and conclusions

Data analysis consisted of inspecting and visualizing the ranges of absorption level values of each wavelength from minimum values (e.g. minM1) to maximum values (e.g. maxM1). For groups M1 and M3, this is shown in Figures 1 and 6 in the wavelength range 1500-1800 and 2700-3000 respectively, while for groups M1 and M6, it is shown in Figures 3 and 8. From these graphs, the difference between these ranges was calculated in order to find the wavelength ranges that clearly distinguish two groups under consideration. Visualization of the difference of the ranges is shown in Figures 4 and 9. These two graphs were obtained for groups M1 and M6 in the wavelength ranges 1500-1800 and 2700-3000. Comparing groups M1 and M3 no such graphs can be created. The final step of the analysis was to determine the significance of individual features (wavelengths) in terms of distinguishing groups. This significance was calculated using the Random Forest algorithm, which allows to assess the influence of individual attributes on the

quality of classification. Significance for individual groups and two wavelength ranges is presented in Figures 2, 5, 7 and 10.

The presented graphs allow us to identify the wavelength ranges that have the greatest influence on the identification of patient groups. In the range 1500-1800 for groups M1 and M3, the ranges of absorption values cannot be separated (Figure 1). The range of values for M1 is practically within the range of M3, which translates into poorer classification quality for both groups. However, feature (wavelength) significance analysis in the context of distinguishing M1 from M3 indicates 6 significant wavelength ranges, see Table 1. Similarly, for distinguishing M1 and M6 groups, 4 interesting significant wavelength ranges can be identified. The intervals in both cases are somewhat similar, but in the case of the distinction between groups M1 and M6 they are much more clearly defined (see Figure 5). This can also be seen in the value range analysis graph for groups M1 and M6 (see Figure 3). One can see clear wave ranges in which the absorption value ranges are separable. By calculating the absorption difference we have obtained Figure 4 which shows these wavelength intervals.

Analogous conclusions are made when analyzing data obtained for the wavelength range 2700-3000. The distinction between M1 and M3 groups, taking into account the space of absorption values, is not so obvious (see Figure 6). This makes it difficult to determine the relevant wavelength intervals in Figure 7, but two distinct intervals can be identified there (Table 1). In contrast, distinguishing the M1 and M6 groups is much easier because the absorption value spaces are largely disjoint (see Figure 8). Hence, the feature significance intervals are very clearly defined (see Figure 9 and 10 and Table 1).

Table 1. Regions of wavenumbers can be used as potential markers to discriminate M1 and M3, and M1 and M6 groups of patients by IR spectrum.

Wavelength spectrum (cm^{-1})	Regions that distinguish M1 and M3 groups (wavenumbers, cm^{-1})	Regions that distinguish M1 and M6 groups (wavenumbers, cm^{-1})
1500-1800	1646-1652 1660-1681 1689-1695 1702-1712 1727-1762 1776-1789	1500-1542 1598-1616 1683-1716 1768-1799
2700-3000	2848-2859 2904-2964	2721-2846 2861-2904 2946-3019

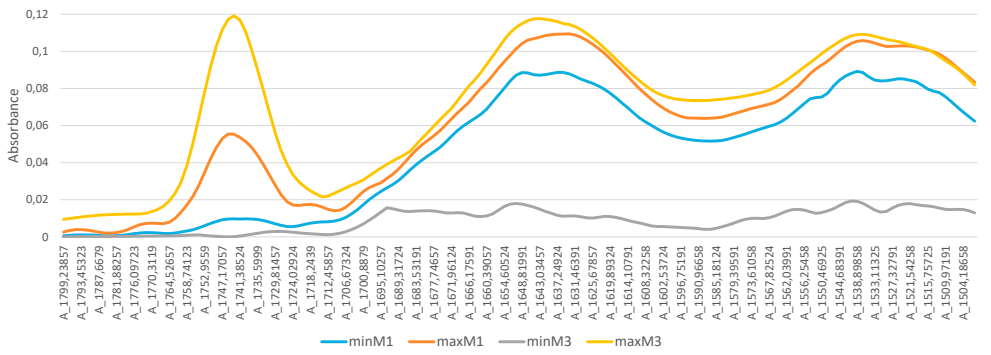

Fig. 1. Range of minimum and maximum wavelength absorption values in the 1500-1800 cm^{-1} wavelength range for patient groups M1 and M3.

The analysis of classification quality (accuracy) and related parameters: sensitivity, precision, and specificity, which are known measures for assessing the accuracy of case diagnoses, are presented in Table 2. The table includes the

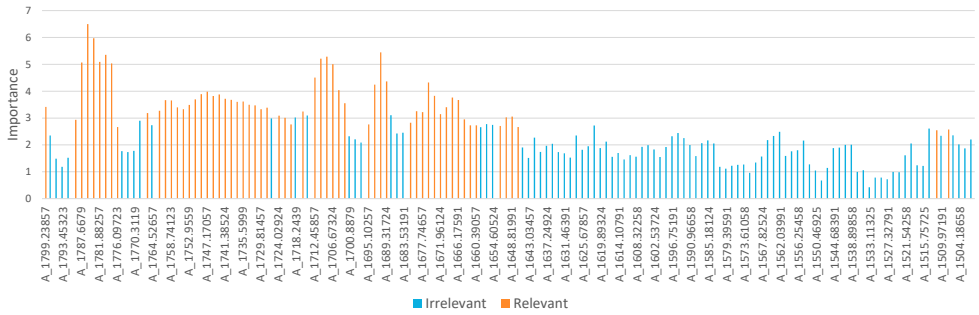

Fig. 2. Mean importance value of attributes (wavelengths) in the wavelength range 1500-1800 cm^{-1} for patient group M1 and M3.

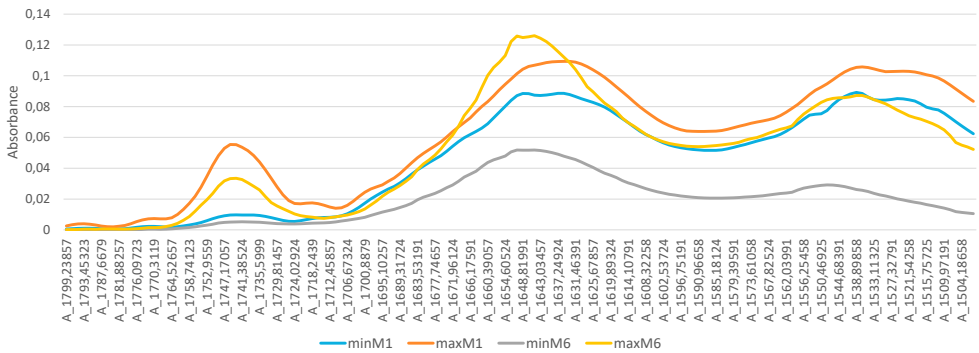

Fig. 3. Range of minimum and maximum wavelength absorption values in the 1500-1800 cm^{-1} wavelength range for patient groups M1 and M6.

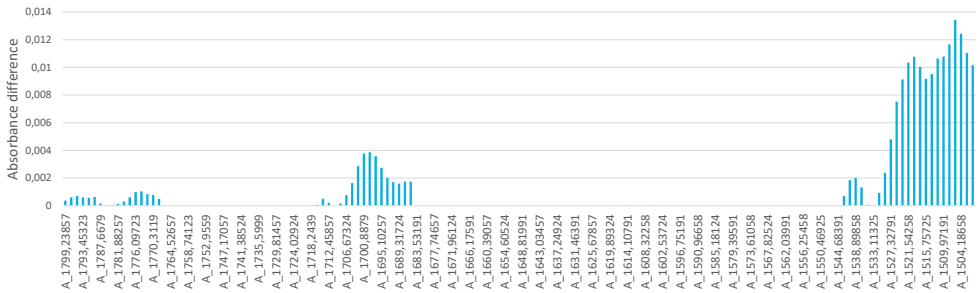

Fig. 4. Difference between absorption levels in the wavelength range 1500-1800 cm^{-1} for patient group M1 and M6.

results obtained using six machine learning algorithms. The obtained values indicate that it is somewhat difficult for the algorithms to distinguish between M1 and M3 groups. The classification quality for the full wavelength range (156 wavelengths, from 1500 to 1800cm^{-1}) ranges from 67.86% to 96.43% depending on the algorithm. Similarly, for the 2700 to 3000cm^{-1} range, accuracy ranges from 75.00% to 96.43%. A similar case can be observed for other

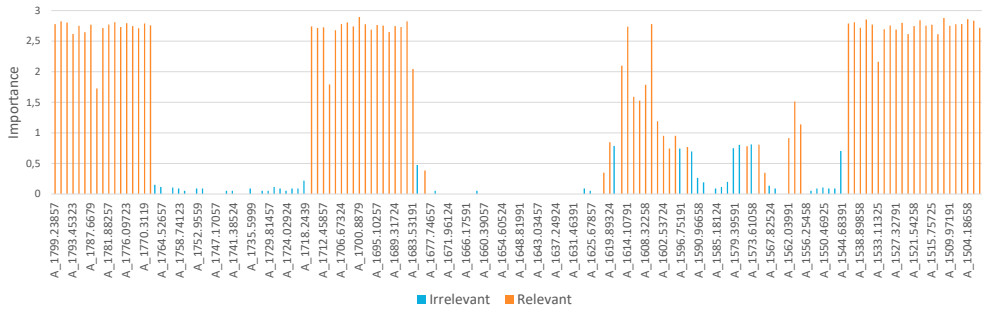

Fig. 5. Mean importance value of attributes (wavelengths) in the wavelength range 1500-1800 cm^{-1} for patient group M1 and M6.

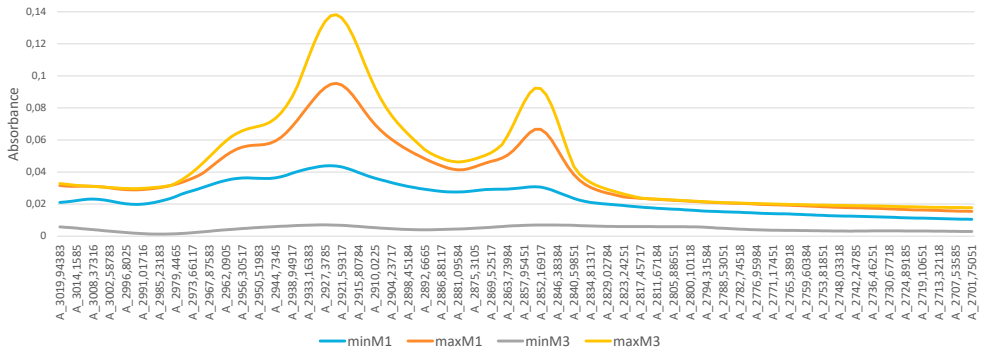

Fig. 6. Range of minimum and maximum wavelength absorption values in the 2700-3000 cm^{-1} wavelength range for patient groups M1 and M3.

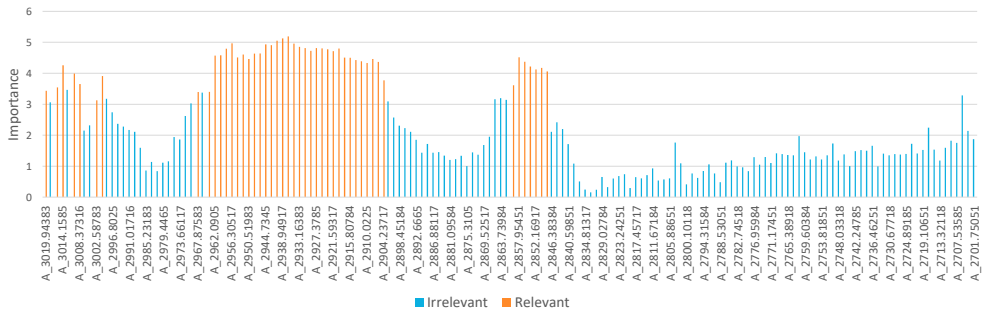

Fig. 7. Mean importance value of attributes (wavelengths) in the wavelength range 2700-3000 cm^{-1} for patient group M1 and M3.

parameters whose values are mostly above 0.9, but for DNN and SVM models they are slightly worse. On the other hand, in the context of distinguishing M1 and M6 groups, the obtained results of classification quality and other parameters are almost perfect. Most models seamlessly identify patients after 6 months of infection from patients

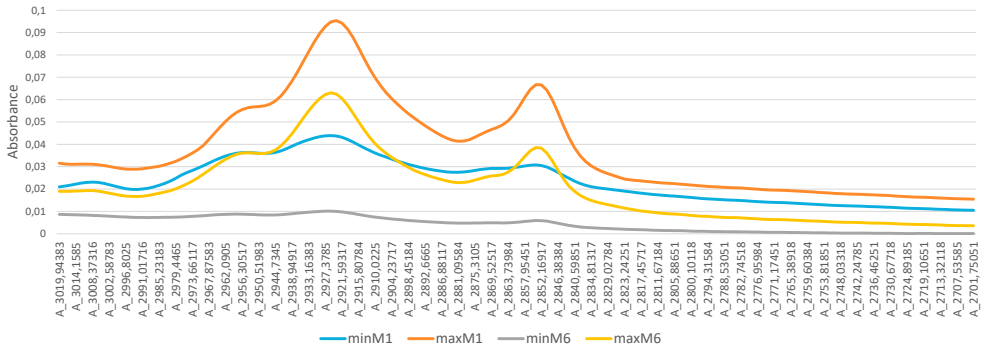

Fig. 8. Range of minimum and maximum wavelength absorption values in the 2700-3000 cm^{-1} wavelength range for patient groups M1 and M6.

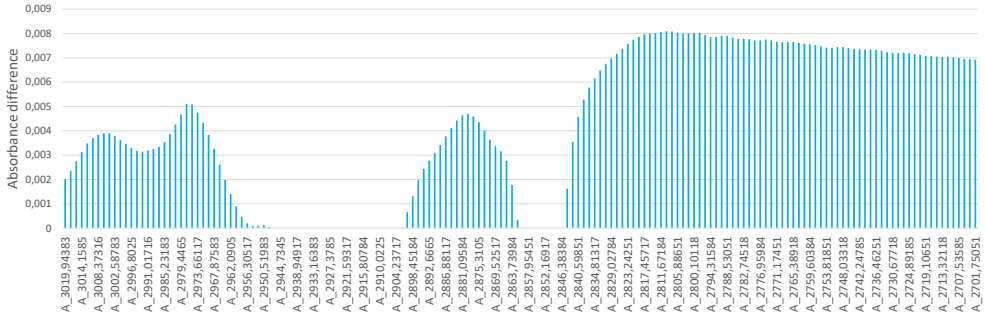

Fig. 9. Difference between absorption levels in the wavelength range 2700-3000 cm^{-1} for patient group M1 and M6.

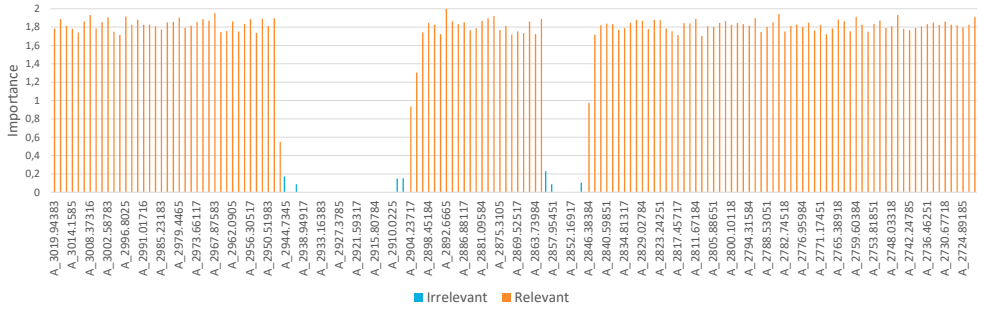

Fig. 10. Mean importance value of attributes (wavelengths) in the wavelength range 2700-3000 cm^{-1} for patient group M1 and M6.

after the first month. Accuracy is 100% while the other parameters are 1. This can be expected from the previous analysis of the value space and wavelet absorption differences for two groups.

As part of the research described in [1], feature (wavelength) selection was also performed using the Random Forest algorithm. As a result of the selection, the wavelengths having the highest relevance in terms of distinguishing

groups were selected. This selection is based on the evaluation of the influence of removing a particular feature from the set on the decrease in classification quality of the random forest model. The original number of wavelengths, i.e., 156 for the range 1500-1800cm^{-1} and 166 for the range 2700-3000cm^{-1}, for groups M1 and M3, decreased to 61 and 47 respectively, and after further analysis and selection of intervals from Table 1, it decreased to 53 and 39. At the same time the quality parameters of classification did not change or turned out to be better. On the other hand, for groups M1 and M6, the number of significant wavelengths decreased from the original to 78 and 138 and to 53 and 21 after further analysis and selection of intervals from Table 1. At the same time, the quality parameters remained ideal.

Table 2. Classification results of group of patients M1 with M3 and M1 with M6 (two-class dataset) with additional classification quality parameters in the range of wavenumbers: *Range 1* from 1500 to 1800cm^{-1}; *Range 2* from 2700 to 3000cm^{-1} using six different machine learning algorithms with original and selected relevant sets of wavenumbers (Table 1).

Dataset	Model	Accuracy		Precision		Sensitivity		Specificity	
		Range 1	Range 2	Range 1	Range 2	Range 1	Range 2	Range 1	Range 2
Groups M1 and M3	RF	96.43	96.43	0.98	0.95	0.95	0.97	0.98	0.96
original features	C5.0	91.67	91.67	0.95	0.93	0.88	0.90	0.95	0.93
(156 for range 1,	kNN (k-3)	96.43	91.67	1	0.98	0.93	0.87	1	0.97
166 for range 2)	DNN	67.86	75.00	0.62	0.73	0.85	0.75	0.52	0.75
	XGBoost	95.24	91.67	0.95	0.90	0.95	0.925	0.96	0.91
	SVM	70.24	76.19	0.64	0.69	0.88	0.90	0.55	0.64
Groups M1 and M3	RF	95.24	95.24	0.98	0.95	0.93	0.95	0.98	0.96
selected features	C5.0	91.67	92.86	0.95	0.93	0.88	0.93	0.95	0.93
(61 for range 1,	kNN (k-3)	97.62	95.24	1	1	0.95	0.91	1	1
47 for range 2)	DNN	67.86	58.33	0.63	0.59	0.8	0.4	0.57	0.75
	XGBoost	95.24	90.48	0.95	0.88	0.95	0.93	0.96	0.89
	SVM	71.43	77.38	0.63	0.71	0.95	0.88	0.5	0.68
Groups M1 and M3	RF	95.24	97.62	0.98	0.98	0.93	0.98	0.98	0.98
selected features	C5.0	91.67	90.48	0.95	0.89	0.89	0.93	0.95	0.88
(53 for range 1,	kNN (k-3)	98.81	95.24	1	1	0.98	0.91	1	1
39 for range 2)	DNN	65.48	63.10	0.61	0.64	0.78	0.53	0.55	0.73
	XGBoost	96.43	90.48	0.95	0.88	0.98	0.93	0.95	0.89
	SVM	72.62	77.38	0.64	0.71	0.98	0.88	0.5	0.68
Groups M1 and M6	RF	100	100	1	1	1	1	1	1
original features	C5.0	98.77	98.77	1	1	0.98	0.98	1	1
(156 for range 1,	kNN (k-3)	100	100	1	1	1	1	1	1
166 for range 2)	DNN	100	100	1	1	1	1	1	1
	XGBoost	100	100	1	1	1	1	1	1
	SVM	100	100	1	1	1	1	1	1
Groups M1 and M6	RF	100	100	1	1	1	1	1	1
selected features	C5.0	98.77	98.77	1	1	0.98	0.98	1	1
(78 for range 1,	kNN (k-3)	100	100	1	1	1	1	1	1
138 for range 2)	DNN	100	100	1	1	1	1	1	1
	XGBoost	100	100	1	1	1	1	1	1
	SVM	98.77	100	1	1	0.98	1	1	1
Groups M1 and M6	RF	100	100	1	1	1	1	1	1
selected features	C5.0	100	100	1	1	1	1	1	1
(53 for range 1,	kNN (k-3)	98.77	100	1	1	0.98	1	1	1
21 for range 2)	DNN	100	100	1	1	1	1	1	1
	XGBoost	100	100	1	1	1	1	1	1
	SVM	100	100	1	1	1	1	1	1

References

[1] Guleken, Z., Tuyji Tok, Y., Jakubczyk, P., Paja, W., Pancerz, K., Shpotyuk, Y., Cebulski, J., Depciuch, J. (2022) "Development of novel spectroscopic and machine learning methods for the measurement of periodic changes in COVID-19 antibody level, *Measurement*, Volume **196**, 111258, *Elsevier*, DOI: 10.1016/j.measurement.2022.111258.

[2] Guleken, Z., Jakubczyk, P., Paja, W., Pancerz, K., Bulut, H., Öten, E., Depciuch, J., Tarhan, N., (2022) "Characterization of Covid-19 infected pregnant women sera using laboratory indexes, vibrational spectroscopy, and machine learning classifications", Talanta, 237(122916)2022, https://doi.org/10.1016/j.talanta.2021.122916.

[3] Ker, J., Wang, L., Rao, J., Lim, T., (2018) "Deep Learning Applications in Medical Image Analysis", *IEEE Access*, vol. **6**, pp. 9375-9389.

[4] Paja, W. (2015) "Medical Diagnosis Support and Accuracy Improvement by Application of Total Scoring from Feature Selection Approach", Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (FEDCSIS 2015) *Annals of Computer Science and Information Systems*, pp. 281-286.

[5] Pancerz, K., Paja, W., Sarzyński, J., Gomuła, J. (2018) "Determining Importance of Ranges of MMPI Scales Using Fuzzification and Relevant Attribute Selection", *Procedia Computer Science* **126**, Elsevier, pp. 2065-2074.

[6] Pati, J., (2019) "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach", *IEEE Access*, vol. **7**, pp. 4232-4238.

[7] Cudek, P., Paja, W., Wrzesień, M. (2011) "Automatic System for Classification of Melanocytic Skin Lesions Based on Images Recognition" in T. Czachórski, S. Kozielski, U. Stanczyk (eds) *Man-Machine Interactions 2*, Advances in Intelligent and Soft Computing, (AISC 103), Springer-Verlag Berlin Heidelberg, pp. 189-196.

[8] Remeseiro, B., Bolon-Canedo, V., (2019) "A Review of Feature Selection Methods in Medical Applications. *Comput Biol Med.* 2019 Sep;**112**:103375.

[9] C. Stoean, R. Stoean, M. Hotoleanu, D. Iliescu, C. Patru and R. Nagy, (2021) "An assessment of the usefulness of image pre-processing for the classification of first trimester fetal heart ultrasound using convolutional neural networks", *2021 25th International Conference on System Theory, Control and Computing (ICSTCC)*, pp. 242-248, doi: 10.1109/ICSTCC52150.2021.9606852.

[10] Syeda-Mahmood, T., (2018) "Role of Big Data and Machine Learning in Diagnostic Decision Support in Radiology", *Journal of the American College of Radiology*, Vol. **15**, Issue 3, pp. 569-576.

[11] Wosiak, A., Kowalski, R., (2020) "Automated Feature Selection for Obstructive Sleep Apnea Syndrome Diagnosis", *Procedia Computer Science* **176**, Elsevier, pp. 1430-1439.

[12] Wosiak, A., Zakrzewska, D. (2018) "Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis" in Czarnowski I. (ed.) *Overcoming "Big Data" Barriers in Machine Learning Techniques for the Real-Life Applications*, *Complexity*, Vol. 2018, Hindawi.

[13] David S. Hui et al., (2020) "The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health – The latest 2019 novel coronavirus outbreak in Wuhan, China, *International Journal of Infectious Diseases*, **91**, 2020, pp. 264–266, DOI: 10.1016/j.ijid.2020.01.009.

[14] Ying-Hui Jin et al., (2020) "A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version)", *Military Medical Research*, **7** (1), 2020, pp. 4, DOI: 10.1186/s40779-020-0233-6.

[15] Quinlan, J.R., C4.5: Programs for Machine Learning", San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[16] Breiman, L., "Random forests", Mach. Learn. **45** (2001) 5–32. https://doi.org/10.1023/A:1010933404324.

[17] Goodfellow, I., Bengio, Y., Courville, (2016) A., "Deep Learning", MIT Press.

[18] Altman, N.S., "An introduction to kernel and nearest-neighbor nonparametric regression", Am. Stat. **46** (1992) 175–185. https://doi.org/10.1080/00031305.1992.10475879.

[19] Chen, T., Guestrin, C., (2016) "XGBoost: A Scalable Tree Boosting System", Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (n.d.). **5** 785-794. https://doi.org/10.1145/2939672

[20] Vapnik, V. Naumovich, (1998) "Statistical learning theory", John Wiley Sons, **736**.

[21] Paja W. (2016) Feature Selection Methods Based on Decision Rule and Tree Models. In: Czarnowski I., Caballero A., Howlett R., Jain L. (eds) Intelligent Decision Technologies 2016. Smart Innovation, Systems and Technologies, vol 57. Springer, Cham. https://doi.org/10.1007/978-3-319-39627-9_6