Erciyes Med J 2022; 44(6): 544–8 • DOI: 10.14744/etd.2022.00868
ORIGINAL ARTICLE – OPEN ACCESS

Prediction of COVID-19 Based on Genomic Biomarkers of Metagenomic Next-Generation Sequencing Data Using Artificial Intelligence Technology

Sami Akbulut[1,2,3], Fatma Hilal Yağın[3], Cemil Çolak[3]

ABSTRACT

Objective: The primary aim of this study was to use metagenomic next-generation sequencing (mNGS) data to identify coronavirus 2019 (COVID-19)-related biomarker genes and to construct a machine learning model that could successfully differentiate patients with COVID-19 from healthy controls.

Materials and Methods: The mNGS dataset used in the study demonstrated expression of 15,979 genes in the upper airway in 234 patients who were COVID-19 negative and COVID-19 positive. The Boruta method was used to select qualitative biomarker genes associated with COVID-19. Random forest (RF), gradient boosting tree (GBT), and multi-layer perceptron (MLP) models were used to predict COVID-19 based on the selected biomarker genes.

Results: The MLP (0.936) model outperformed the GBT (0.851), and RF (0.809) models in predicting COVID-19. The three most important biomarker candidate genes associated with COVID-19 were IFI27, TPTI, and FAM83A.

Conclusion: The proposed model (MLP) was able to predict COVID-19 successfully. The results showed that the generated model and selected biomarker candidate genes can be used as diagnostic models for clinical testing or potential therapeutic targets and vaccine design.

Keywords: Artificial intelligence, Boruta, COVID-19 pandemic, feature selection, multi-layer perceptron, SARS-CoV-2 virus

Cite this article as:
Akbulut S, Yağın FH, Çolak C. Prediction of COVID-19 Based on Genomic Biomarkers of Metagenomic Next-Generation Sequencing Data Using Artificial Intelligence Technology. Erciyes Med J 2022; 44(6): 544-8.

[1]Department of Surgery, İnönü University Faculty of Medicine, Malatya, Türkiye
[2]Department of Public Health, İnönü University Faculty of Medicine, Malatya, Türkiye
[3]Department of Biostatistics and Medical Informatics, İnönü University Faculty of Medicine, Malatya, Türkiye

Submitted
23.04.2022

Revised
05.05.2022

Accepted
13.05.2022

Available Online
22.06.2022

Correspondence
Sami Akbulut,
İnönü University Faculty of Medicine, Department of Surgery, Malatya, Türkiye
Phone: +90 422 341 06 60
e-mail: akbulutsami@gmail.com

INTRODUCTION

The SARS-CoV-2 virus, which first emerged in China in December 2019, triggered the world's first coronavirus pandemic. The disease caused by the virus, coronavirus 2019 (COVID-19), spread quickly around the world and became a significant public health problem (1, 2). More than 500 million confirmed cases and more than 6 million deaths have been reported globally, according to the World Health Organization (2). The rapid transmission of this novel virus and the sudden rise in the number of patients and patient deaths led to substantial disruption and prompted immediate research efforts.

Common COVID-19 symptoms include fever, cough, pneumonia, weakness, chest pressure, diarrhea, and shortness of breath. Because the symptoms are similar to those of influenza, it can be difficult to diagnose COVID-19 in the early stage. However, it is critical to identify and isolate positive cases as soon as possible to reduce transmission of the virus and begin treatment. As a result of the quick disease progression and a high mortality rate, healthcare resources in many countries became severely overburdened. It was critical to determine factors that might assist in diagnosis and influence the prognosis and treatment methods (3, 4).

One of the most common methods of detecting COVID-19 is reverse transmission polymerase chain reaction (RT-PCR) testing. However, the sensitivity and accuracy of RT-PCR testing has been called into question in various investigations. RT-PCR tests can produce a significant number of false positives and negatives (4, 5). Computed tomography (CT) scans, chest X-rays, and ultrasound scans are also used to detect COVID-19. The literature includes research based on machine learning (ML) that has successfully identified COVID-19 utilizing these images. However, there are also certain drawbacks. COVID-19 and other types of viral pneumonia may have few visible characteristics in the early stage, and it can be difficult to distinguish between respiratory viruses, so medical scan images may be insufficient (6). A lack of findings of COVID-19 on chest X-rays, CT scans, or ultrasound scans is not proof of absence. At the time of the outbreak, there were also few annotated images of different types that could be used in image-based studies (6, 7).

Although numerous approaches to detect the virus have been examined, thorough knowledge of the scientific and genetic character of the virus is extremely valuable. The genomic structure, gene regions, protein binding sites, attachment, and neutralizing structures of the virus, among other information, provides important guidance. In addition, detailed research of the viral host response aids understanding of how it spreads and what treatments might work best (8, 9).

The risk of pre-symptomatic transmission and developing severe symptoms increases when diagnosis is delayed (10). Though progress has been made, a reliable method of early detection and diagnosis of COVID-19 infection continues to be important.

Metagenomic next-generation sequencing (mNGS) is an agnostic method to identify microorganisms from a clinical specimen using high-throughput sequencing and automated bioinformatic analysis. Coinfections can affect illness development and prognosis, and mNGS can provide valuable information on the composition of the microbiome. Therefore, mNGS technology can provide more precise information and serve as a useful tool to detect the coronavirus responsible for COVID-19 and other organisms that may have an effect on the prognosis (11).

The interest in artificial intelligence (AI) and ML technologies has increased rapidly in recent years. AI and ML are data-driven and use statistical concepts to learn from examples and errors and to improve with additional data. AI uses logic to simulate human reasoning. ML is a subfield of AI. Models are used to promote development based on experience without explicit programming. AI and ML can be used as a clinical support system, including the identification of disease-related biomarkers, aiding in disease diagnosis and prognosis, estimating treatment efficacy, refining individual treatment plans, and acting as an early warning system. The use of computer algorithms can advance early detection and treatment of critical illnesses, and may facilitate management of high-risk patients (12–14).

ML has been used with various medical datasets, such as clinical, radiological image/video, genetic information, and proteins, in the effort to combat COVID-19 (5, 15–17). ML methods can help to distinguish individuals infected with COVID-19 from healthy individuals using genomic data and to help find new treatment options. The goal of this study was to use mNGS data to find COVID-19-related biomarker candidate genes and to build an ML model that can predict COVID-19.

MATERIALS and METHODS

Dataset

This research used an open mNGS dataset of 234 patients provided by the University of California. In the cohort, 141 (60.3%) of the patients had negative PCR results for COVID-19 and 93 (39.7%) had positive PCR results. Differential expression analysis revealed expression of 15,979 genes in the upper airway (18).

Data Preprocessing and Development of Predictive Models

The Boruta method was used to select biomarker genes associated with COVID-19. This process iteratively removes variables that have been statistically proven to be less relevant to the response (COVID-19 in the current study) (19). Afterward, the new dataset was divided with 80% to be used for training and 20% as a test set. This split was randomly repeated 50 times, and the mean scores were calculated to evaluate the models. Three models were constructed to predict COVID-19: random forest (RF), gradient boosting tree (GBT), and multi-layer perceptron (MLP) models.

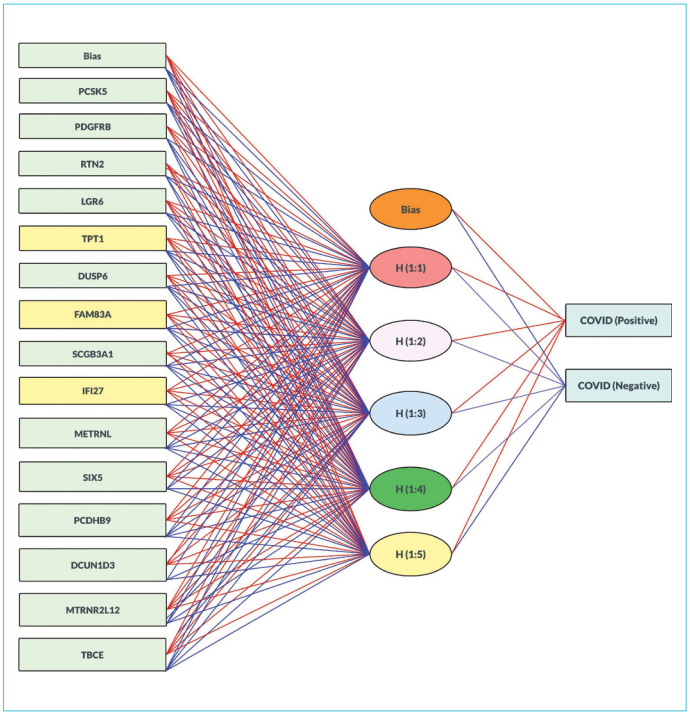

Figure 1. Schematic representation of the multi-layer perceptron network created in this study

The RF method can be used as a classification algorithm to make predictions based on numerous decision trees. The cumulative result of the trees provides a reasonable prediction. The model also identifies the most significant variables that explain the dependent variable, which frequently leads to improved performance (20, 21). In this study, 100 trees were used in the RF model.

The GBT method combines multiple decision trees to make a single powerful learner. All trees are connected in a series and each tree tries to reduce the error of the previous tree as much as possible. Gradient algorithms are often slow to learn from data because of this sequential connection, but GBT outperforms classical ML approaches (22).

An MLP is a type of neural network that is used to supplement feed-forward neural networks. The input layer receives the signal to be processed. The output layer carries out tasks, such as prediction and classification. The actual computational engine of the MLP is comprised of an arbitrary number of hidden layers that are sandwiched between the input and output levels of the MLP. Data flow in the forward direction from the input layer to the output layer is like a feed-forward network. Backpropagation learning, is used to train the neurons in the MLP. These models can approximate any continuous function and solve problems that are not linearly separable. Pattern categorization, identification, and prediction are among the most common applications of MLP (23, 24). The MLP network created in this study is presented in Figure 1.

The performance of the models created was evaluated according to accuracy, F1-score, precision, recall, and area under the receiver operating characteristic curve (AUC) criteria and the results were compared.

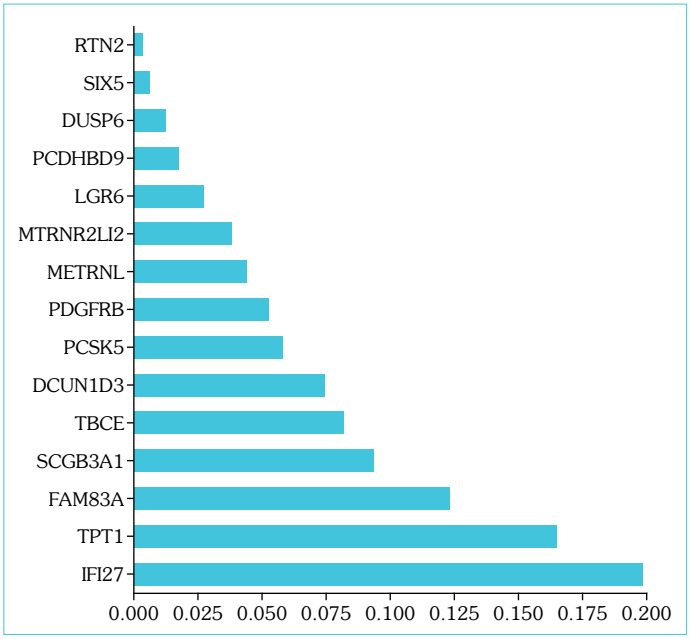

Figure 2. Importance plot of genes associated with coronavirus 2019 generated in this study

Statistical Analyses

Qualitative variables were summarized as numbers and percentages, and a chi-squared test was used to analyze the data. Quantitative variables were reported as the median and interquartile range (IQR). Two groups were compared using the Mann-Whitney U test. A p value of <0.05 was considered significant. The Cohen d effect size was calculated for variables with a significant p value. The effect size was interpreted as a small at 0.20–0.50, medium at 0.50–0.80, and large at >0.80 in Mann-Whitney U tests (25). All of the statistical analyses were performed using Python 3.9 (Python Software Foundation, Fredericksburg, VA, USA) and IBM SPSS Statistics for Windows, Version 26.0 (IBM Corp., Armonk, NY, USA).

RESULTS

Demographic descriptive statistics of the patients are provided in Table 1. The median age of the patients with a negative COVID-19 PCR test result was 54.5 years (IQR: 28.5 years); 74 (52.48%) of these patients were female and 67 (47.52%) were male. The median age of the patients with a positive COVID-19 PCR test result was 44 years (IQR: 28 years); 50 (53.76%) of these patients were female and 43 (46.24%) were male. The Boruta feature selection method yielded 15 genes associated with COVID-19 from the 15,979 genes identified in the mNGS dataset. Descriptive statistics for the 15 biomarker candidate genes are given in Table 2.

There were statistically significant differences between the patient groups that were positive and negative for COVID-19 in the PDGFRB, RTN2, TPT1, DUSP6, SCGB3A1, METRNL, PCDHB9, DCUN1D3, PCSK5, LGR6, FAM83A, and IFI27 genes (p<0.05). The PCSK5, LGR6, FAM83A, and IFI27 genes revealed a higher expression level in the COVID-19-positive group; expression of PDGFRB, RTN2, TPT1, DUSP6, SCGB3A1, METRNL, PCDHB9, and DCUN1D3 was lower. The effect size results indicated that the IFI27, TPT1, and FAM83A genes had the greatest clinical impact for COVID-19.

Table 1. Descriptive statistics of the demographic characteristics

	Groups		p
	COVID-19 negative	COVID-19 positive	
Gender, n (%)			
Female	74 (52.48)	50 (53.76)	0.848*
Male	67 (47.52)	43 (46.24)	
Age, years, median (IQR)	54.5 (28.5)	44 (28)	<0.001**

*: Chi-squared test; **: Mann-Whitney U test; COVID-19: Coronavirus 2019; IQR: Interquartile range

Table 2. Descriptive statistics and effect size of biomarker candidate genes associated with COVID-19

Genes	Groups		p*	Effect size
	COVID-19 negative Median (IQR)	COVID-19 positive Median (IQR)		
PCSK5	311 (845)	891 (2305)	<0.001	0.613
PDGFRB	11 (62)	2 (13)	<0.001	0.61
RTN2	32 (68)	15 (33)	0.002	0.439
LGR6	18 (26)	50 (62)	<0.001	0.937
TPT1	3438 (7193)	2801 (2854)	<0.001	0.968
DUSP6	175 (338)	102 (119)	<0.001	0.487
FAM83A	656 (1368)	1700 (2488)	<0.001	0.943
SCGB3A1	30 (89)	13 (30)	0.012	0.352
IFI27	254 (465)	1014 (1337)	<0.001	1.069
METRNL	148 (278)	80 (96)	<0.001	0.608
SIX5	35 (67)	28 (47)	0.227	NS
PCDHB9	9 (20)	3 (7)	<0.001	0.506
DCUN1D3	314 (605)	117 (113)	<0.001	0.899
MTRNR2L12	18 (106)	15 (33)	0.282	NS
TBCE	371 (487)	280 (972)	0.548	NS

*: Mann-Whitney U test; COVID-19: Coronavirus 2019; IQR: Interquartile range; NS: Not significant

Table 3 shows the performance criteria used to evaluate the test data set of the RF, GBT, and MLP models. The accuracy, F1-score, precision, recall, and AUC values obtained from the RF model for the prediction of COVID-19 were 80.9%, 78%, 72.7%, 84.2%, and 91.7%, respectively. The accuracy, F1-score, precision, recall, and AUC values of the GBT model were 85.1%, 82.1%, 8%, 84.2%, and 92.1%, respectively. Finally, the results of the MLP model were accuracy: 93.6%, F1-score: 92.7%, precision: 86.4%, recall: 100%, AUC: 97.7%. The MLP model outperformed the RF and GBT models for COVID-19 prediction. The importance coefficients of genes thought to be associated with COVID-19 are given in Table 4 and illustrated in a graph in Figure 2.

Table 3. Performance measures for models built to predict COVID-19

Models	Accuracy	F1-score	Precision	Recall	AUC
RF	0.809	0.78	0.727	0.842	0.917
GBT	0.851	0.821	0.800	0.842	0.921
MLP	0.936	0.927	0.864	1.000	0.977

AUC: Area under the receiver operating characteristic curve; GBT: Gradient boosting tree; MLP: Multi-layer perceptron; RF: Random forest

The IFI27, TPTI, and FAM83A genes explained 19.88%, 16.53%, and 12.34% of predicted COVID-19, respectively. These were the three most important biomarker candidate genes associated with COVID-19 (Fig. 2).

DISCUSSION

The effects of COVID-19 did not follow the pattern of many other infectious diseases, and wealthier countries with greater health resources were more afflicted than low-income countries with fewer health resources. The consequences included broad economic and social disruption.

Thorough identification of the virus's source and genome structure are imperative to understand and prevent further spread and mutation, develop future therapeutic targets, and to ensure treatment effectiveness. As disease-causing host genetic features are identified, new strategies for addressing this new virus can be proposed.

Comprehensive analysis of genomic biomarkers will enable more effective large-scale screening and early diagnosis of COVID-19. Initial analyses that have identified candidate gene biomarkers for COVID-19 must be supplemented with studies to isolate specific biomarkers suitable for clinical use. The aim of the current study was to identify some specific qualitative host biomarkers associated with COVID-19 infection using an mNGS dataset to contribute to improved clinical diagnosis of COVID-19 and the development of new drugs and vaccines.

Analysis of the dataset using the Boruta feature selection method resulted in the genes IFI27, TPT1, FAM83A, SCGB3A1, TBCE, DCUN1D3, PCSK5, PDGFRB, METRNL, MTRNR2LI2, LGR6, PCDHBD9, DUSP6, SIX5, and RTN2 as biomarker candidate genes for COVID-19. Shojaei et al. (26) found that IFI27 transcription was an early predictor for COVID-19 outcomes. They reported that IFI27 was expressed in the respiratory tract of COVID-19 patients and that high IFI27 expression was associated with the presence of a high viral load. They also observed that the systemic host response as measured by IFI27 expression was associated with COVID-19 severity. Our findings indicated that the IFI27 gene was the most important gene associated with COVID-19. In another study, it was emphasized that the TPT1 gene had an important role in the development of COVID-19 (27). We also found that the TPT1 gene may be a biomarker candidate. Zhang et al. (28) reported that the FAM83A and LGR6 genes were associated with the SARS-CoV-2 virus. These were also among the genes identified in our study.

Table 4. Importance coefficients of the genes used as input in the MLP model

Feature	Feature importance
IFI27	0.1988
TPTI	0.1653
FAM83A	0.1234
SCGB3A1	0.09394
TBCE	0.08234
DCUN1D3	0.07485
PCSK5	0.05798
PDGFRB	0.0529
METRNL	0.0444
MTRNR2LI2	0.03839
LGR6	0.02735
PCDHBD9	0.018
DUSP6	0.01266
SIX5	0.00628
RTN2	0.0034

MLP: Multi-layer perceptron

Three ML models (RF, GBT, and MLP) were used with a dataset containing selected biomarker genes to examine the ability to distinguish between healthy controls and patients with COVID-19 infection. The MLP method performed better than the RF and GBT methods in predicting COVID-19. The performance evaluation revealed accuracy, F1-score, precision, recall, and AUC values of 93.6%, 92.7%, 86.4%, 100%, and 97.7%, respectively for the MLP method. We found that the IFI27 (19.88%), TPT1 (16.53%), and FAM83A (12.34%) genes were the three most important genes for predicting COVID-19.

CONCLUSION

The selected qualitative biomarker genes are associated with COVID-19 and can contribute to distinguishing COVID-19 cases from healthy controls. The application of ML can effectively help identify potential diagnostic biomarkers and candidate drug targets and help establish a standardized workflow for relevant analyses.

Acknowledgements: We would like to commend all health care professionals who were always in the frontline. They took the courage and responsibility of treating all patients during these hard times and despite risking their own lives. They are the real heroes.

Ethics Committee Approval: The İnönü University Non-Interventional Clinical Research Ethics Committee granted approval for this study (number: 2022/ 3670).

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – FHY, SA; Supervision – CÇ; Resource – FHY, SA; Materials – FHY, SA; Data Collection and/or Processing – FHY, SA; Analysis and/or Interpretation – FHY, SA, CÇ; Literature Search – FHY, SA; Writing – FHY, CÇ; Critical Reviews – FHY, SA, CÇ.

Conflict of Interest: The authors have no conflict of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES

1. Sahin TT, Akbulut S, Yilmaz S. COVID-19 pandemic: Its impact on liver disease and liver transplantation. World J Gastroenterol 2020; 26(22): 2987–99. [CrossRef]
2. Myoung J. Two years of COVID-19 pandemic: where are we now?. J Microbiol 2022; 60(3): 235–7. [CrossRef]
3. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet 2020; 395(10223): 497–506. [CrossRef]
4. Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nature machine intelligence 2020; 2(5): 283–8. [CrossRef]
5. Yagin FH, Güldoğan E, Ucuzal H, Colak C. A computer-assisted diagnosis tool for classifying COVID-19 based on Chest X-Ray images. Konuralp Med J 2021; 13(S1): 438–45. [CrossRef]
6. Arentz M, Yim E, Klaff L, Lokhandwala S, Riedo FX, Chong M, et al. Characteristics and outcomes of 21 critically ill patients with COVID-19 in Washington State. JAMA 2020; 323(16): 1612–4. [CrossRef]
7. Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. IEEE Rev Biomed Eng 2021; 14: 4–15.
8. Hoque MN, Rahman MS, Ahmed R, Hossain MS, Islam MS, Islam T, et al. Diversity and genomic determinants of the microbiomes associated with COVID-19 and non-COVID respiratory diseases. Gene Rep 2021; 23: 101200. [CrossRef]
9. Zhang Y, Pan Y, Zhao X, Shi W, Chen Z, Zhang S, et al. Genomic characterization of SARS-CoV-2 identified in a reemerging COVID-19 outbreak in Beijing's Xinfadi market in 2020. Biosaf Health 2020; 2(4): 202–5. [CrossRef]
10. Ding W, Nayak J, Swapnarekha H, Abraham A, Naik B, Pelusi D. Fusion of intelligent learning for COVID-19: A state-of-the-art review and analysis on real medical data. Neurocomputing 2021; 457: 40–66. [CrossRef]
11. Mostafa HH, Fissel JA, Fanelli B, Bergman Y, Gniazdowski V, Dadlani M, et al. metagenomic next-generation sequencing of nasopharyngeal specimens collected from confirmed and suspect COVID-19 patients. mBio 2020; 11(6): e01969–20. [CrossRef]
12. Perçin İ, Yağin FH, Güldoğan E, Yoloğlu S. ARM: An interactive web software for association rules mining and an application in medicine. 2019 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE; 2019. pp.1–5. [CrossRef]
13. Bellini V, Valente M, Bertorelli G, Pifferi B, Craca M, Mordonini M, et al. Machine learning in perioperative medicine: a systematic review. J Anesthesia, Analgesia and Critical Care 2022; 2(1): 1–13. [CrossRef]
14. Yagin FH, Yagin B, Arslan AK, Colak C. Comparison of performances of associative classification methods for cervical cancer prediction: Observational study. Turkiye Klinikleri J Biostatistics 2021; 13(3): 266–72. [CrossRef]
15. Shiri I, Salimi Y, Pakbin M, Hajianfar G, Avval AH, Sanaat A, et al. COVID-19 prognostic modeling using CT radiomic features and machine learning algorithms: Analysis of a multi-institutional dataset of 14,339 patients. Comput Biol Med 2022; 145: 105467. [CrossRef]
16. Abubaker Bagabir S, Ibrahim NK, Abubaker Bagabir H, Hashem Ateeq R. Covid-19 and Artificial Intelligence: Genome sequencing, drug development and vaccine discovery. J Infect Public Health 2022; 15(2): 289–96. [CrossRef]
17. Suvarna K, Biswas D, Pai MGJ, Acharjee A, Bankar R, Palanivel V, et al. Proteomics and machine learning approaches reveal a set of prognostic markers for COVID-19 severity with drug repurposing potential. Front Physiol 2021; 12: 652799. [CrossRef]
18. Mick E, Kamm J, Pisco AO, Ratnasiri K, Babik JM, Castañeda Get al. Upper airway gene expression reveals suppressed immune responses to SARS-CoV-2 compared with other respiratory viruses. Nat Commun 2020; 11(1): 5854. [CrossRef]
19. Li Z, Guo W, Ding S, Feng K, Lu L, Huang T, et al. Detecting blood methylation signatures in response to childhood cancer radiotherapy via machine learning methods. Biology (Basel) 2022; 11(4): 607.
20. He S, Wu J, Wang D, He X. Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. Chemosphere 2022; 290: 133388. [CrossRef]
21. Palimkar P, Shaw RN, Ghosh A. Machine learning technique to prognosis diabetes disease: random forest classifier approach. Advanced Computing and Intelligent Technologies. Springer; 2022.p.219–44.
22. Larestani A, Mousavi SP, Hadavimoghaddam F, Hemmati-Sarapardeh A. Predicting formation damage of oil fields due to mineral scaling during water-flooding operations: Gradient boosting decision tree and cascade-forward back-propagation network. J Petroleum Scien Engineer 2022; 208(A): 109315. [CrossRef]
23. Kruse R, Mostaghim S, Borgelt C, Braune C, Steinbrecher M. Multilayer perceptrons. Computational intelligence. Springer; 2022.p.53–124. [CrossRef]
24. Xu Y, Li F, Asgari A. Prediction and optimization of heating and cooling loads in a residential building based on multi-layer perceptron neural network and different optimization algorithms. Energy 2022; 240: 122692. [CrossRef]
25. Cohen J. The world is round (p<. 05). American Psychologist 1994; 49(12): 997–1003. [CrossRef]
26. Shojaei M, Shamshirian A, Monkman J, Grice L, Tran M, Tan CW, et al. IFI27 transcription is an early predictor for COVID-19 outcomes; a multi-cohort observational study. medRxiv 2021. Available from: URL: https://www.medrxiv.org/content/10.1101/2021.10.29.21265555v1. [CrossRef]
27. Hasankhani A, Bahrami A, Sheybani N, Aria B, Hemati B, Fatehi F, et al. Differential co-expression network analysis reveals key hub-high traffic genes as potential therapeutic targets for COVID-19 pandemic. Front Immunol 2021; 12: 789317. [CrossRef]
28. Zhang YH, Li H, Zeng T, Chen L, Li Z, Huang T, et al. Identifying transcriptomic signatures and rules for SARS-CoV-2 infection. Front Cell Dev Biol 2021; 8: 627302. [CrossRef]