Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2025;14(1):134-40

Determination of risk factors for COVID-19 patients using the CatBoost machine learning technique

Turkan Mutlu Yar[1], Ipek Balikci Cicek[2], Tuba Gul[3], Ulku Karaman[1]

[1]Ordu University, Faculty of Medicine, Department of Medical Parasitology, Ordu, Türkiye
[2]İnönü University, Faculty of Medicine, Department of Biostatistics, Malatya, Türkiye
[3]Ordu University, Faculty of Medicine, Department of Neurology, Ordu, Türkiye

Received 11 December 2024; Accepted 06 January 2025
Available online 01 March 2025 with doi: 10.5455/medscience.2024.12.165

Abstract

A comprehensive understanding of the risk factors related to coronavirus disease 2019 (COVID-19), which can cause various severe clinical pictures, is vital for predicting the progression of the disease and administering the appropriate treatment to prevent adverse outcomes. Therefore, this study aimed to predict the presence of COVID-19 and to identify possible risk factors for COVID-19 patients using the CatBoost machine learning (ML) technique. CatBoost was used to classify COVID-19. The model results were evaluated using sensitivity (Sen), specificity (Spe), negative predictive value (NPV), positive predictive value (PPV), accuracy (Acc), F1score, precision, recall, and area under the curve (AUC) metrics. In the modeling phase, a10 - fold cross-validation approach was employed. Lastly, variable importance values were derived through modeling. The dataset used in the article is open access and was collected in India in May 2020 this data is available at https://www.kaggle.com/code/anirbansarkar823/covid-presence-detection-ensembling. When the results of The CatBoost model successfully classified the modeling dataset, with Acc, Sen, Spe, PPV, NPV, F1score, recall, precision, and AUC values of 97.79%, 97.21%, 98.55%, 93.23%, 98.77%, 98.66%, 96.00%, 97.79%, and 99.62%, respectively. Considering the findings show that the risk factors related to COVID-19 infection, which has become a threat to humanity, can be successfully determined. Similar studies can update information on the transmission routes of the disease and positively change its course.

Keywords: Diarrhea, machine learning, random forest, bagged CART, CatBoost model

Introduction

Coronaviruses belong to the family Coronaviridae, subfamily Ortho coronavirinae, which includes many subgenera. Viruses of these subgenera can be found in humans and other domestic and wild animals, such as bats, pigs, cats, dogs, rodents, and poultry [1]. Coronaviruses are single-stranded RNA positive-sense viruses that affect humans' and other animals' upper and lower respiratory tracts. They have an average diameter of 60 to140 nm and appear to have a crown-likes hape under an electron microscope [2]. While Human coronaviruses cause 5 % 10% of upper respiratory tract infections in adults and can lead to top neumonia, particularly in older people [3]. our types of coronaviruses (229E, HKU1, NL63, and OC43) cause illness in humans, usually in the form of a common cold. Apart from these, severe acute respiratory syndrome coronavirus (SARS-CoV; first detected in 2002) and MERS-CoV (first detected in 2012) can cause serious respiratory diseases and no so comial epidemics [4]. Moreover, a novel coronavirus was discovered in a study on the reporting of viral pneumonia cases of unclear etiology in Wuhan, China, in December 2019. The virus was initially dubbed 2019-nCoV but was later named SARS-CoV-2 because of its resemblance to SARS-CoV. The disease that it causes, called coronavirus disease 2019 (COVID-19), quickly spread around the world, and the World Health Organization (WHO) declar edit a pandemic in March 2020 [5,6]. By February 28, 2022, the WHO had reported 434,154,739 official cases and 5,944,342 deaths related to COVID-19 (WHO, 2022).

CITATION
Yar TM, Balikci Cicek O, Karaman U, Gul T. Determination of risk factors for COVID-19 patients using the CatBoost machine learning technique. Med Science. 2025;14(1):134-40.

Corresponding Author: Turkan Mutlu Yar, Ordu University, Faculty of Medicine, Department of Medical Parasitology, Ordu, Türkiye
Email: turkanmutluyar@odu.edu.tr

DOI: 10.5455/medscience.2024.12.165

In suspected cases, COVID-19 is often diagnosed by detecting SARS-CoV-2 in respiratory tract samples using real-time reverse transcription polymerase chain reaction (RT-PCR) [7].

COVID-19 can be asymptomatic or may cause a variety of clinical manifestations, ranging from mild upper respiratory tract infection to severe pneumonia, thrombo embolic complications, and multi-organ failure [8]. The associated morbidity and mortality are largely caused by acute respiratory failure syndrome (ARDS) due to viral pneumonia. A comprehensive understanding of the immune pathology determining the severity of the disease and the risk factors linked associated with the disease is necessary for clinicians to identify individuals who require priority therapy to prevent disease progression and severe consequences [9]. Moreover, the early identification of risk factors can help identify critically ill patients, administer the appropriate treatment, and prevent mortality [10].

Recent advances in medical applications based on artificial intelligence (AI) have enabled intelligent healthcare systems to perform multiple crucial functions, including early disease prediction and diagnosis, clinical decision support, patient support, prognosis, and mortality risk assessment for severe illnesses [11]. These developments have provided physicians with valuable support. One of the most significant AI techniques is machine learning (ML). ML algorithms use historical data to replicate human ways of perceiving and learning [12]. CatBoost is an ML model based on gradient boosting that was proposed by Prokhoren kova et al. in 2017 [13]. The CatBoost algorithm creates decision trees (DTs) symmetrically, keeping the tree structure and depth under control and producing faster and higher results. Unlike other gradient boosting–based algorithms, it uses categorical features directly hence its name, which is derived from ıt is “category” and “boosting” [14].

His study aimed to predict the presence of COVID-19 and to identify possible risk factors for COVID-19 patients using the CatBoost technique.

Material and Methods

Dataset and Variables

The dataset used in this study was obtained from the web page https://www.kaggle.com/code/anirbansarkar823/covid-presence-detection-ensembling/notebook. The data set used in the article is open access and was collected in India in May 2020.

The data set included 1,051 (19.3%) non-COVID-19 patients and 4,383 (80.7%) patients with COVID-19. The variables included in the dataset are shown in Table 1.

Due to the use of a dataset in the study, ethical committee approval is not required. The document number indicating that Ethical Committee of Non-Invasive Scientific Researches, Ordu University approval is not required is: 08.09.2023/225.

Table 1. Variables included in the data set

Variable	Categories
Breathing problems	No
	Yes
Fever	No
	Yes
Dry cough	No
	Yes
Sore throat	No
	Yes
Asthma	No
	Yes
Chroniclung disease	No
	Yes
Headache	No
	Yes
Heart disease	No
	Yes
Diabetes	No
	Yes
Hypertension	No
	Yes
Fatigue	No
	Yes
Gastrointestinal	No
	Yes
Contact with a COVID-19 patient	No
	Yes
Traveling abroad	No
	Yes
Visiting public places	No
	Yes
Attending a large gathering	No
	Yes
Family working in public places	No
	Yes
Sanitization from market	No
	Yes
Wearing a mask	No
	Yes
COVID-19	No
	Yes

Categorical boosting (CatBoost)

Gradient-boosted DT (GBDT) techniques are ensemble approaches that build a few DT stobe employed routinely for classification or regression [15]. This study employed CatBoost for model training and evaluation. CatBoost is an advanced GBDT model based on a sophisticated gradient descent–based

DOI: 10.5455/medscience.2024.12.165

ensemble learning approach. A sequence of DTs is formed one after another through model training, with each succeeding tree demonstrating a diminishing loss. In other words, each DT learns from the previous tree to enhance the model's mance, creating a robust learning capability. CatBoost differs from ordinary GBDT in two important aspects: category information–handling efficiency and ordered boosting [16]. At the training stage, learning classifiers effectively handle numeric data, but categorical properties are difficult to decode. Because of this, categorical features are converted to meaning fulinformation using traditional techniques, such as one-hotencoding or gradient statistics [17]. The former method utilizes binary values to replace each feature category , whereas the latter uses gradient statistics to substitute for the original categorical feature at each boosting step to estimate a value. However, for category qualities with a high degree of repeatability, these techniques require considerable memory and processing resources. To overcome this problem, CatBoost employs efficient, modified target-based statistics to effectively maintain category characteristics throughout the training phase, saving considerable computing time and resources [18].

Another crucial aspect is the order denature of the CatBoost algorithm's boosting process. After multiple boosting steps, conventional GBDT suse all training data to develop aprediction model. This strategy shifts the model's predictions, resulting in a unique target leak age issue. Utilizing an ordered boosting architecture, CatBoost can overcome this difficulty. Moreover, unlike conventional learning classifiers, CatBoost employs a large number of training dataset permutations to prevent over fitting [19].

Biostatistical Analysis

Qualitative data were expressed as number sand percentages. Statistically significant differences between the target variable and input variables were assessed using, where appropriate, Pearson's chi-squared test and the continuity correction test. P-values of<0.05 were considered statistically significant. All analyses were performed using IBMSPSS Statistics 26.0.

Machine Learning Modeling and Performance Evaluation

The data set was divided into training and testing data sets in an 80:20 ratio. An n-fold cross-validation approach was used to assess the model's validity. In n-fold cross-validation, the data are divided into subsamples, one of which is used for testing, while the remaining n −1 subsamples are used to train the model. In this study, 10-fold cross-validation was performed. Accuracy (Acc), sensitivity (Sen), specificity (Spe), positive predictive value (PPV), negative predictive value (NPV), F1score,recall, precision, and area under the curve (AUC) metrics were used to evaluate the model's performance. Furthermore, variable significance was established to determine the degree to which the input variables contributed to the output variable. Python 3.10 was used for modeling [20].

Results

Table 2 shows the distribution of the variables included in the dataset.

Table 2. Distribution of the variables in the dataset

Variable	Categories	Count	Percentage (%)
Breathing problems	**No**	1,814	33.4
	Yes	3,620	66.6
Fever	**No**	1,161	21.4
	Yes	4,273	78.6
Dry cough	**No**	1,127	20.7
	Yes	4,307	79.3
Sore throat	**No**	1,481	27.3
	Yes	3,953	72.7
Rhinorrhea	**No**	2,482	45.7
	Yes	2,952	54.3
Asthma	**No**	2,920	53.7
	Yes	2,514	46.3
Chronic lung disease	**No**	2,869	52.8
	Yes	2,565	47.2
Headache	**No**	2,698	49.7
	Yes	2,736	50.3
Heart disease	**No**	2,911	53.6
	Yes	2,523	46.4
Diabetes	**No**	2,846	52.4
	Yes	2,588	47.6
Hypertension	**No**	2,771	51.0
	Yes	2,663	49.0
Fatigue	**No**	2,613	48.1
	Yes	2,821	51.9
Gastrointestinal	**No**	2,883	53.1
	Yes	2,551	46.9
Traveling a broad	**No**	2,983	54.9
	Yes	2,451	45.1
Contact with a COVID-19 patient	**No**	2,708	49.8
	Yes	2,726	50.2
	Yes	2,510	46.2
Visiting public places	**No**	2,614	48.1
	Yes	2,820	51.9
Family working in public places	**No**	3,172	58.4
	Yes	2,262	41.6
Wearing a mask	**No**	5,434	100.0
	Yes	0	0.0
Sanitization from market	**No**	5,434	100.0
	Yes	0	0.0
COVID-19	**No**	1,051	19.3
	Yes	4,383	80.7

DOI: 10.5455/medscience.2024.12.165

There sults of the statistical analysis of the independent variables interm softhetar get variable are shown in Table 3.

As shown in Table 3, there were statistically significant differences between the target variable the (COVID-19) no/yes categories in terms of fever, breathing problems, sore throat, dry cough, asthma, hypertension (HT), headache, chronic lung disease, heart disease, diabetes mellitus (DM), fatigue, traveling abroad, contact with a COVID-19 patient, attending a large gathering, visiting public places, and family members working in public places ($p<0.05$).

Table 3. results of the statistical analysis of the relationships between the target variable and the independent variables

Variable		COVID-19		P-value
		No	Yes	
Breathing problems	**No**	800 (76.12)	1,014 (23.13)	<0.001*
	Yes	251 (23.88)	3,369 (76.87)	
Fever	**No**	535 (50.90)	626 (14.28)	<0.001*
	Yes	516 (49.10)	3,757 (85.72)	
Dry cough	**No**	622 (59.18)	505 (11.52)	<0.001*
	Yes	429 (40.82)	3,878 (88.48)	
Sore throat	**No**	767 (72.98)	714 (16.29)	<0.001*
	Yes	284 (27.02)	3,669 (83.71)	
Rhinorrhea	**No**	474 (45.10)	2,008 (45.81)	0.677*
	Yes	577 (54.90)	2,375 (54.19)	
Asthma	**No**	661 (62.89)	2,259 (51.54)	<0.001*
	Yes	390 (37.11)	2,124 (48.46)	
Chronic lung disease	**No**	494 (47.00)	2,375 (54.19)	<0.001**
	Yes	557 (53.00)	2,008 (45.81)	
Headache	**No**	492 (46.81)	2,206 (50.33)	0.040*
	Yes	559 (53.19)	2,177 (49.67)	
Heart disease	**No**	592 (56.33)	2,319 (52.91)	0.046*
	Yes	459 (43.67)	2,064 (47.09)	
Diabetes	**No**	594 (56.52)	2,252 (51.38)	0.003*
	Yes	457 (43.48)	2,131 (48.62)	
Hypertension	**No**	646 (61.47)	2,125 (48.48)	<0.001*
	Yes	405 (38.53)	2,258 (51.52)	
Fatigue	**No**	458 (43.58)	2,155 (49.17)	0.001*
Gastrointestinal	**No**	554 (52.71)	2,329 (53.14)	0.804*
	Yes	497 (47.29)	2,054 (46.86)	
Traveling abroad	**No**	1,051 (100.00)	1,932 (44.08)	<0.001**
	Yes	0 (0.00)	2,451 (55.92)	
Contact with a COVID-19 patient	**No**	907 (86.30)	1,801 (41.09)	<0.001*
	Yes	144 (13.70)	2,582 (58.91)	
Attending a large gathering	**No**	983 (93.53)	1,941 (44.28)	<0.001*
	Yes	68 (6.47)	2,442 (55.72)	
Visiting public places	**No**	634 (60.32)	1,980 (45.17)	<0.001*
	Yes	417 (39.68)	2,403 (54.83)	
Family working in public places	**No**	783 (74.50)	2,389 (54.51)	<0.001*
	Yes	268 (25.50)	1,994 (45.49)	
Wearing a mask	**No**	1051 (100.00)	4,383 (100.00)	–
	Yes	0 (0.00)	0 (0.00)	
Sanitization from market	**No**	1051 (100.00)	4,383 (100.00)	–
	Yes	0 (0.00)	0 (0.00)	

*Pearson's chi-squared test; **Continuity correction test

DOI: 10.5455/medscience.2024.12.165

The classification matrix for the CatBoost model used to classify the data set is shown in Figure 1.

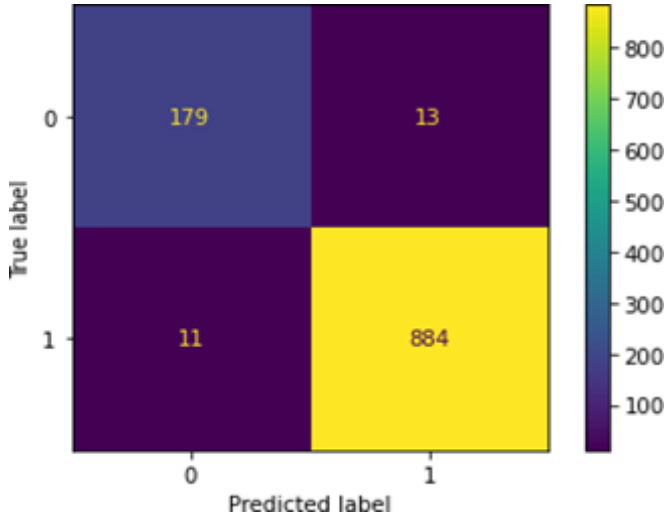

Figure 1. Classification matrix for the CatBoost model

The model's performance metrics used at the testing stage are shown in Table 4.

Table 4. CatBoost performance metrics at the testing stage

Metric	Value (%)
Acc	97.79
Sen	97.21
Spe	98.55
PPV	93.23
NPV	98.77
F1score	98.66
Recall	96.00
Precision	97.79
AUC	99.62

Acc: accuracy, Sen: sensitivity, Spe: specificity, PPV: positive predictive value, NPV: negative predictive value, AUC: area under the curve

As shown in Figure 2, the CatBoost model's Acc, Sen, Spe, PPV, NPV, F1 score, recall, precision, and AUC values were 97.79%, 97.21%, 98.55%, 93.23%, 98.77%, 98.66%, 96.00%, 97.79%, and 99.62%, respectively.

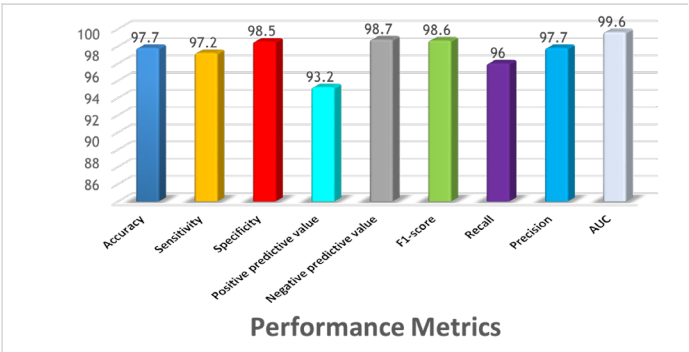

Figure 2. The importance values of the variables obtained as a result of the CatBoost model

The variables' importance values are shown in Table 5. Traveling abroad, attending a large gathering, breathing problems, dry cough, sore throat, fever, contact with a COVID-19 patient, heart disease, family members working in public places, chronic lung disease, fatigue, headache, and asthma were the most important variables associated with COVID-19.

Table 5. Importance values of the independent variables based on the CatBoost model

Variable	Importance value
Traveling abroad	100
Attending a large gathering	89.69
Breathing problems	87.42
Dry cough	83.17
Sore throat	61.13
Fever	50.42
Contact with a COVID-19 patient	25.5
Heart disease	18.53
Family working in public places	14.62
Chronic lung disease	10.88
Fatigue	10.08
Headache	8.73
Asthma	6.4

Discussion

The COVID-19 pandemic infected over 45 million individuals and killed over 1.2 million people globally. The clinical presentations of the disease are broad, ranging from no symptoms to severe disease and death. Severe and critical cases accounted for 14% and 5% of COVID-19 patients, respectively. This placed heavy burden on healthcare systems, consuming the bulk of medical resources and contributing to the majority of fatalities [21]. The average incubation period of COVID-19 is five to six days. Symptoms may appear within 2-14 days after infection, and contagiousness may begin one to two days before symptoms appear. Transmission from person to person direct occur by contact with mucous membranes or inhalation of infected droplets. Fecal and intrauterine modes of transmission have also been reported [4]. Thus, transmission occurs through close contact not only with patients and their infected interests but also with infected family members and/or friends during the incubation period [22].

Uncertainties about the pathogenesis of COVID-19 and insufficient knowledge of its course and complications led to difficulties in the follow-up of patients, especially at the beginning of the pandemic. The accurate identification of patients, especially those with a severe prognosis requiring hospitalization or even intensive care, was one of the main problems when the case load increased. In to make the most efficient use of strained healthcare systems. It was necessary to clearly identify high-risk patients [23,24]. Therefore, identifying risk factors for severe disease and the need for intensive care became vitally important for the correct management of COVID-19 patients and the pandemic.

DOI: 10.5455/medscience.2024.12.165

ML models are a branch of AI that can be used in a wide variety of fields. Their main purpose is to determine effective factors and the relationships between them, based on which they can make predictions [25]. Thus, this study aimed to predict COVID-19 and to identify the risk factors associated with the disease in order of importance using CatBoost, a machine learning (ML) model from gradient boosting algorithms. The model's performance analysis obtained Acc, Sen, Spe, PPV, NPV, F1score, recall, precision, and AUC values of 97.79%, 97.21%, 98.55%, 93.23%, 98.77%, 98.66%, 96.00%, 97.79%, and 99.62%, respectively, indicating that the model accurately predicted COVID-19 The variable importance tained as a result of the model analysis showed that the variables most closely associated with the disease were in order of importance, traveling abroad, attending large a gathering, breathing problems, dry cough, sore throat, fever, contact with a COVID-19 patient, heart disease, family members working in public places, chronic lung disease, fatigue, headache, and asthma.

In A previous study found that traveling abroad was an important risk factor for COVID-19, since the disease began to spread in Europe and some Central Asian countries before it was detected in our country. In that study, nine (56.2%) patients had a history of traveling to a European country, and seven (43.8%) had a history of returning from hope [4]. These findings are in line with the results of this study.

HT, cardiovascular disease (CVD), DM, immunosuppressive therapy or disease, chronic lung disease, organ transplantation, obesity, chronic kidney failure, and smoking are among the risk factors for death [26] A meta-analysis that included seven studies including 1.576 patients found that the most common comorbidities were HT, DM, CVD, and respiratory diseases. [27]. Another meta-analysis of six studies conducted in China that involved a total of 1,558 patients identified chronic obstructive pulmonary disease (COPD), CVD, DM, and HT as the most important independent risk factors. In this meta-analysis, six studies were included and there were a total of 1.558 COVID-19 patients. [28]. Another study suggested that advanced age and comorbidities were associated with severe disease, with HT (42%), DM (21%), CHF (10%), allergic asthma (7%), COPD (6%), coronary artery disease (CAD) (2%), depression (1%), and epilepsy (1%) being the most common comorbidities [4]. In A retrospective study of 1,590 cases from 575 hospitals in China found that 3.7% of the patients were found to had CVD [29]. In a study of 5,700 hospitalized COVID-19patients in NewYork, which accounted for more than 30% of the cases in the USA, 595 (11.1%) patients had CAD, and 371 (6.9%) had congestive heart failure (CHF) [30]. In a meta-analysis, a total of 34 studies were included, with a total of 6.263 COVID-19 cases, including 1.727 and 4.536 severe and non-severe patients, respectively. This study showed that the risk of severe disease was three to four times higher in patients with CVD [31]. Another meta-analysis involving a total of 1.527 COVID-19 patients and a total of six studies found that the relative risk of CVD (RR=3.30, 95% CI) was higher among those who needed intensive care than among those who did not require intensive care follow-up [32]. In this study, heart disease, chronic lung disease, headache, and asthma were important risk factors for COVID-19.

The most prevalent symptoms of COVID-19 infection are fever, cough, fatigue, and dyspnea [27]. In A study involving 41 patients found that the most frequently reported symptoms were muscular ache or weariness (44%), cough (76%), and fever (98%) [7]. In another study, the most frequently reported symptoms were fever (88.7%), cough (67.8%), exhaustion (38.1%), sputum (33.4%), breathing problems (18.6%), cough (13.9%), and headache (13.6%). Another study reported cough (93%), fever (42%), dyspnea (22%), fatigue (8%), sore throat (7%), and headache (5%) as the most prevalent symptoms [4]. In this study, some of the most significant risk factors for COVID-19 were dry cough, breathing problems, fever, sore throat, and fatigue.

Conclusion

In conclusion, it is vital to broaden our understanding of COVID-19 through further research on the related risk factors, modes of transmission, and clinical characteristics of this disease, which has become a threat to humanity. This study makes a valuable contribution to the literature and may change the course of COVID-19 by determining the risk factors and clinical features associated with COVID-19 infection.

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical Approval
Due to the use of a dataset in the study, ethical committee approval is not required. The document number indicating that Ethical Committee of Non-Invasive Scientific Researches, Ordu University approval is not required is: 08.09.2023/225.

References

1. Ministry of TS.COVID-19 (SARS-CoV2infection) guidance. https://covid19.saglik.gov.tr/TR-66301/covid-19-rehberi.html access date: 16.07.2020.
2. GuoY-R, Cao Q-D, Hong Z-S, et al.The origin, transmission and clinical therapies oncoronavirus disease 2019 (COVID-19) outbreak–an update on the status. Mil Med Res. 2020;7:11.
3. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med. 2020;382:1199-207.
4. Günal Ö, Türe E, Bayburtlu M, et al. Evaluation of patients diagnosed with COVID-19 in terms of risk factors. Microbiol Bul. 2020;54:575-82.
5. Hui DS, Azhar EI, Madani TA, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. Int J Infect Dis. 2020;91:264-6.
6. Lu R, Zhao X, Li J, et al. Genomic character isationand epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395:565-74.

7. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus inWuhan, China. Lancet. 2020;395:497-506. Erratum in: Lancet. 2020;395:496.
8. Baj J, Juchnowicz-Karakula H, Teresinski G. et al.COVID-19: Specific and non-specific clinical manifestations and symptoms: the current state of knowledge. J Clin Med. 2020;9:1753.
9. Sokolowska M, Lukasik ZM, Agache I, et al. Immunologyof COVID-19: mechanisms, clinical outcome, diagnostics, and perspectives—are port of the European Academy of Allergy and Clinical Immunology (EAACI). Allergy. 2020;75:2445-76.
10. Alizadehsani R, Alizadeh Sani Z, Behjati M, et al. Risk factors prediction, clinical outcomes, and mortality in COVID-19 patients. J Med Virol. 2021;93:2307-20.
11. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health (Amst). 2021;20:100178.
12. Aldweesh A, Derhab A, Emam AZ. Deep learning approaches for anomaly-based intrusion detection systems: a survey, taxonomy, and open issues. Knowledge-Based Systems. 2020;189:105124.
13. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv preprint. 2018;arXiv:181011363.
14. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data. 2020;7:94.
15. Friedman JH. Stochastic gradient boosting. Computational Statistics &Data Analysis. 2002;38:367-78.
16. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. Advancesin Neural Information Processing Systems. 2018;31.
17. Ke G, Meng Q, Finley T, et al.Light GBM: a highly efficient gradient boosting decision tree. Advances in neuralin formation processing systems. 2017;30.
18. Hussain S, Mustafa MW, Jumani TA, et al. A novel feature engineered-CatBoost-based supervised machine learning frame work for electricity theft detection. Energy Reports. 2021;7:4425-36.
19. Hancock J, Khoshgoftaar TM. Medicare fraud detection using CatBoost. 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). doi: 10.1109/IRI49571.2020.00022.
20. Inden M. Short introduction to Python 3.10. Python Challenges.Springer; 2022;635-43.
21. Gao YD, Ding M, Dong X, et al. Risk factors for severe and critically ill COVID-19 patients: a review. Allergy. 2021;76:428-55.
22. Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020;382:1708-20.
23. Jain V, Yuan JM. Predictive symptom sand comorbidities for severe COVID-19 and intensive care unit admission: asystematic review and meta-analysis. Int J Public Health. 2020;65:533-46.
24. Qiu P, Zhou Y, Wang F, et al. Clinical characteristics, laboratory outcome characteristics, comorbidities, and complications of related COVID-19 deceased: asystematic review and meta-analysis. Aging Clin Exp Res. 2020;32:1869-78.
25. Yoo I, Alafaireet P, Marinov M, et al. Datamining in health care and biomedicine: a survey of the literature. J Med Syst. 2012;36:2431-48.
26. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult in patients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 2020;395:1054-62. Erratum in: Lancet. 2020;395:1038.
27. Yang J, Zheng Y, Gou X, et al. Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: asystematic review and meta-analysis. Int J Infect Dis. 2020;94:91-5.
28. Wang B, Li R, Lu Z, Huang Y. Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis. Aging (Albany NY). 2020;12:6049-57.
29. Liang WH, Guan WJ, Li CC, et al. Clinical characteristics and outcomes of hospitalized patients with COVID-19 treated in Hubei (epicentre) and outside Hubei (non-epicentre): anation wide analysis of China. Eur Respir J. 2020;55:2000562.
30. Richardson S, Hirsch JS, Narasimhan M, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the NewYork City area. JAMA.2020;323:2052-9. Erratum in: JAMA. 2020 May 26;323:2098.
31. Wang X, Fang X, Cai Z, et al. Comorbid chronic diseases and acute organ injuries are strongly correlated with diseases everity and mortality among COVID-19 patients: a systemic review and meta-analysis. Research (Wash D C). 2020;2020:2402961.
32. Li B, Yang J, Zhao F, et al. Prevalence and impact of cardio vascular metabolic diseases on COVID-19 in China. Clin Res Cardiol. 2020;109:531-8.