International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621

Prediction of Health Issues During Covid-19 using Machine Learning Technique

Payel Ghosh[1], Shubhi Awasthi[2], Dibiya Sarkar[3], Subha Roy[4], Mrs. Sulekha Das[5]
[1]*UG- Computer Science and Engineering, Techno Engineering College Banipur, Habra, Kolkata*
[2]*UG- Computer Science and Engineering, Techno Engineering College Banipur, Habra, Kolkata*
[3]*UG- Computer Science and Engineering, Techno Engineering College Banipur, Habra, Kolkata*
[4]*UG- Computer Science and Engineering, Techno Engineering College Banipur, Habra, Kolkata*
[5]*Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur, Habra, Kolkata*
ORCID ID : 0000-0001-8970-8045

ABSTRACT
A novel deadly virus named COVID-19 was born in China in 2019. In early 2020, the COVID-19 virus spread worldwide, causing severe infections and deaths due to its infectious characteristics and no medical treatment. It has been termed the most consequential global crisis since the World Wars. The defence involved against COVID-19 spread includes sores like social distancing, personal hygiene, wearing a mask, and using sanitizer. The disaster, affecting billions of lives economically and socially improved its infection impact and has motivated the scientific community to come up with solutions based on computerized digital technologies for diagnosis, prevention, rescue, and estimation of COVID-19. An Artificial Intelligence based analysis made an effort to focus on the available data concerning COVID-19. All of these scientific efforts demand that the data brought to service for analysis should be open source to assist the extension, validation, and collaboration of the work in the fight against the global disaster pandemic. We survey and compare research works in these directions that are accompanied by open-source data and code. We hope that the article will provide the scientific community with an initiative to start open-source-extensible and transparent research in the collective fight against the COVID-19 pandemic. The authors collected data on student spending time on online platforms and facing health issues during the lockdown period from DELHI NCR. The authors gather information primarily from the 'UCI Repository'" https://archive.ics.uci.edu/ml/datasets". The dataset for this study is extracted to predict the number of students who faced health issues during the lockdown period for COVID-19. The authors used logistic regression, 50-50,66-34,80-20 train-test splits and 10-fold cross-validation to analyze the data set.
Keywords – COVID-19, Logistic regression, 10-fold cross-validation

INTRODUCTION

The World Health Organization (WHO) declared the COVID-19 virus a deadly pandemic. More than ten million cases and 503862 deaths across the world as per WHO statistics of 30 June 2020 were predicted. COVID-19 is caused by Severe acute respiratory syndrome Coronavirus2 (SARS-CoV-2). It was declared a pandemic by WHO on March 11, 2020. It took several months to cure due to its medical trials on humans. Further, the cure for COVID-19 is delayed due to possible genetic mutations shown by the COVID-19 virus. Billions of people are socially, economically, and medically affected by a pandemic situation with drastic changes in social relationships, health policies, trade, work, and educational environments. The global pandemic calls for immediate action as it was a threat to human society. The pandemic has motivated the research community to provide front-line medical service staff with cutting-edge virus mitigation, detection, and prevention research. No one expected that Covid-19 would turn our world into a disaster and would bring major changes to people's lifestyles. The virus has spread everywhere like a disaster wild forest fire due to which a lot of changes people accepted all over the world, and it took some time for everyone to adjust to a new normal life. Online learning is rapidly becoming one of the most effective ways to communicate

DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621

with the education system. The impact of the virus was so strong that online education became an important part of our growing world, which resulted in the closure of schools and physical appearance in public places, and no further physical interaction of teachers with students.
In this paper, a machine learning method is applied to investigate information regarding health issues of student's during the lockdown, to assess the prescient intensity of these systems. To this aim, a logistic regression model has been developed for health issues and student's spent their time during covid-19 in lockdown period.. This enables the model to achieve better accuracy in the prediction and prognosis stages. In this paper eight independent fields has been selected. In the end, the performance of the proposed algorithm is assessed when it applies to a **COVID-19** database.

LITERATURE STUDY

- Kunal Chaturvediand Nidhi Sing represented a Survey on the experience of online learning among students during COVID-19. They compared their research and found that their proposed method performed better. The best classification accuracy of their survey reached Negative response-38.3% Positive response-28.4%, and Average response-33.4%.
- Aayush Gargre presented a cross-sectional survey conducted with a sample size of 1182 students of different age groups. The best classification accuracy of their survey reached a positive response.
- Duduzile Ndwandwe and Charles S Wiysonge represented a cross-sectional survey conducted with a sample size of 1182 students of different age groups.they showed details about the variation of the maximum and the minimum values of the accuracy with the classification of the survey. The best classification accuracy of their survey reached a positive response.
- Majdi M Sabahelzain, Kenneth Hartigan-Go, and Heidi J Larson represented a Review on – (1)how governments handle the Covid-19 response more broadly,(2)populism,(3) vaccine equity, particularly when it plays out as vaccine nationalism..
- Amir Ahmad representedthe Prediction of confirmed cases.He used the traditional machine learning regress techniquesique to come up with an accurate result. The best classification accuracy of their survey reached Suggestions to predict the number of confirmed cases of Covid-19 were identified successfully.

Kunal Chaturvedi Nidhi Sing	Surveyon the experience of online learning among students duringCOVID-19.	-	Negative response-38.3% Positive response-28.4% Average response-33.4%
AYUSH GARG	a cross-sectional survey is conducted with a sample size of 1182 students of different age groups	-	Positive response
Duduzile Ndwandwe and Charles S Wiysonge	Article on a survey of COVID-19 vaccine	-	Research advances incorporating vaccinology have ensured that the most

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621

			critical public health intervention is developed
Majdi M Sabahelzain, Kenneth Hartigan-Go and Heidi J Larson	Review on – (1)how governments handle the Covid-19 response more broadly, (2)populism, (3) vaccine equity, particularly when it plays out as vaccine nationalism.	-	-
Amir Ahmad	Prediction of confirmed cases	Traditional Machine Learning Regression	Suggestions to predict the number of confirmed cases of Covid-19, identified successfully.

METHODOLOGY

DATA: In this paper, data were taken from the UCI Machine Learning Repository. In the dataset, there is a total of 19 columns. Out of these columns, 8 columns are taken – " Health issue during lockdown" i.e. people dealing with health issues, "Time spent on fitness" means people under the fitness level, "Time spent on sleep" expresses people spent time on sleeping (numeric), "Time spent on social media" provides data on people spend time on social media (numeric), "Number of meals per day" means people receiving meal (numeric) "Time spent on a self-study" expresses people spending time on self-study, "Time spent on Online Class" describes people spent time on online class, "age of the subject," tells us about the age of the total population in a particular area, "Change in your weight" means people receiving change in their weight, "Time spent on TV" shows people spending time on TV. The main motive is to find out about health issues faced by students during COVID-19 through fields. Now one can say, in a growing state if there are most the people affected then the growth of the infected virus will be high. That's why it may play a role in the health issues index. The population is also one of the reasons for health issues because if the population will be very much high then the infected level will increase automatically.

Table 1-For attributes mean and standard deviation:

ATTRIBUTES	DEFINITION	MEAN	STANDARD DEVIATION
Health issues during lockdown	people dealing with health issues	**0.32**	**0.46**
Time spent on fitness	people under the fitness level	**0.74**	**0.72**
Time spent on sleep	people spend time sleeping	**7.87**	**1.61**

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621

Time spent on social media	people spend time on social media	**2.36**	**1.75**
Number of meals per day	people receiving meal	**2.87**	**0.82**
Time spent on self-study	people spending time on self-study	**2.93**	**2.15**
Time spent on online class	people spent time in online classes	**3.07**	**2.09**
Age of subject	age of the total population in a particular area	**2.29**	**5.19**
Change in your weight	Weight of the people	**1.16**	**0.74**
Time spent on TV	People spent time on TV	**0.98**	**1.25**

RESEARCH METHOD: logistic regression (LR) is a type of statistical regression method. It is used to analyse the relationship between a dependent variable and two or more independent variables. This method was taken for this paper because there are more than two independent controlled variables. In this research paper, health issues during lockdown (y) are the main data while the first independent variable is Time spent on fitness (x0), the second independent variable is Time spent on sleep (x1), the third independent variable is Time spent on social media (x2), the fourth independent variable is Number of meals per day(x3), the fifth independent variable is Time spent on the self-study (x4), the fifth independent variable is Time spent on online class(x5), the sixth independent variable is Age of subject(x6), the seventh independent variable isChange in your weight(x7),the eighth independent variable isTime spent on TV(x8).

STEP 1 :

•**LOGISTIC REGRESSION:**We selected the dependent and independent variables from the dataset. After that, we find a linear relationship between dependent and independent variables i.e. LR. We have taken 2/3, ½, and 1/4 of data from the dataset as trained data and the rest of the data as test data. All the trained and test data were taken using a random module.

• **CROSS VALIDATION:** In logistic regression, Cross-validation is a very useful technique. It is a method in which data resampling is done to assess the generalization ability of predictive models and to prevent overfitting. Overfitting is an insidious problem because no one can detect it unless cross-validations are implemented carefully. So cross-validation is one of the most vital steps for verifying the results of this paper correctly. We have done cross-validation by dividing the total dataset into 10 parts and creating 10 sub-datasets using the random module. At first, we are taking the first 9 sub-datasets as trained data and the last one as test data. In the 2nd case, we took the first 8 sub-datasets and the last one as trained data, and after repeating the steps, the 9th no sub-dataset as test d, data, and so on.

• **CONFUSION MATRIX:**A confusion matrix i.e. also called an error matrix, is a type of matrix or a table where we put the performance or the results of the MLR model i.e. the test data. It is the shortest way to see and understand the result of the model. The confusion matrix contains four variables– TP, TN, FP, and FN.

- TP stands for ‘true positive’ which shows the student has health issues and the test is positive.
- TN stands for ‘true negative’ which shows the student does not have health issues and the test is negative.
- FP stands for ‘false positive’ which indicates the student does not have health issues and the test is positive.It is also called a TYPE 1 ERROR.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621

- FN stands for 'false negative' which indicates the student has health issues and the test is negative. It is also called type 2 errors. Grammarly checked-84%

PREDICTED

POSITIVE **NEGATIVE**

•**Accuracy:** Accuracy represents the ratio of the number of times the model can make the correct prediction to the total number of predictions.
•**Sensitivity:** Sensitivity is defined as the ability of a test to correctly identify students with health issues during lockdown in COVID-19.
•**Specificity:** Specificity shows the ability of a test to correctly identify students without health issues during lockdown in COVID-19.
•**Precision:** Precision expresses the method by which one can say how correctly predicted cases turned positive.

STEP 2 - Selecting the suitable model of multiple linear regression: LR is used to find the relationship between the dependent and independent variables. So in this model, we are doing 2/3,½, and ¼ data checking and also cross-validation.

STEP 3 - DEVELOPING EQUATION OF MLR:

In this model, the hypothesis that is used:

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621

H0: None of the controlled variable x0,x1,x2,x3,x4,x5,x6,x7,x8and x9 is significantly related to Y
Ha: At least one of the controlled variables X1, X2 a, and X3, is significantly related to Y
The model of multiple linear regression can be represented as:

$$\mathbf{y = a + b_1x_1 + b_2x_2 + \ldots\ldots\ldots + b_nx_n} \frac{\sum y * \sum x^2 - \sum x * \sum (x * y)}{n\sum x^2 - (\sum x)^2}$$

$$bi = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{where, } 1,2,3 \ldots\ldots n$$

y =Health issue during lockdown =people under the health issues during lockdown (numeric)
b0 =b6= constant variable
b1 = coefficient of a first controlled variable
b2 = coefficient of the second controlled variable
b3 = coefficient of the third controlled variable
b4 = coefficient of the fourth controlled variable and so on
b5= coefficient of the fifth controlled variable and so on
b7= coefficient of the sixth controlled variable and so on
x0 = Time spent on fitness = people spent time on fitness (numeric)
x1 = Time spent on sleep = people spent time on sleep(numeric)
x2 = Time spent on social media = people spent time on social media
x3 = Number of meals per day = people receiving meal
x4=Time spent on self-study
x5= Time spent on online class
x6=Age of subject

LOGISTIC REGRESSION – Sig(x) = $\frac{1}{1+e^{-x}}$
DEVELOPING EQUATION OF CONFUSION MATRIX:
Let's take-
TP= TRUE POSITIVE
TN= TRUE NEGATIVE
FP= FALSE POSITIVE
FN= FALSE NEGATIVE
Now,

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

Sensitivity = $\frac{TP}{TP+FN}$

Specificity = $\frac{TN}{TN+FP}$

Precision = $\frac{TP}{TP+FP}$

Recall = $\frac{TP}{TP+FN}$

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621

FLOW CHART

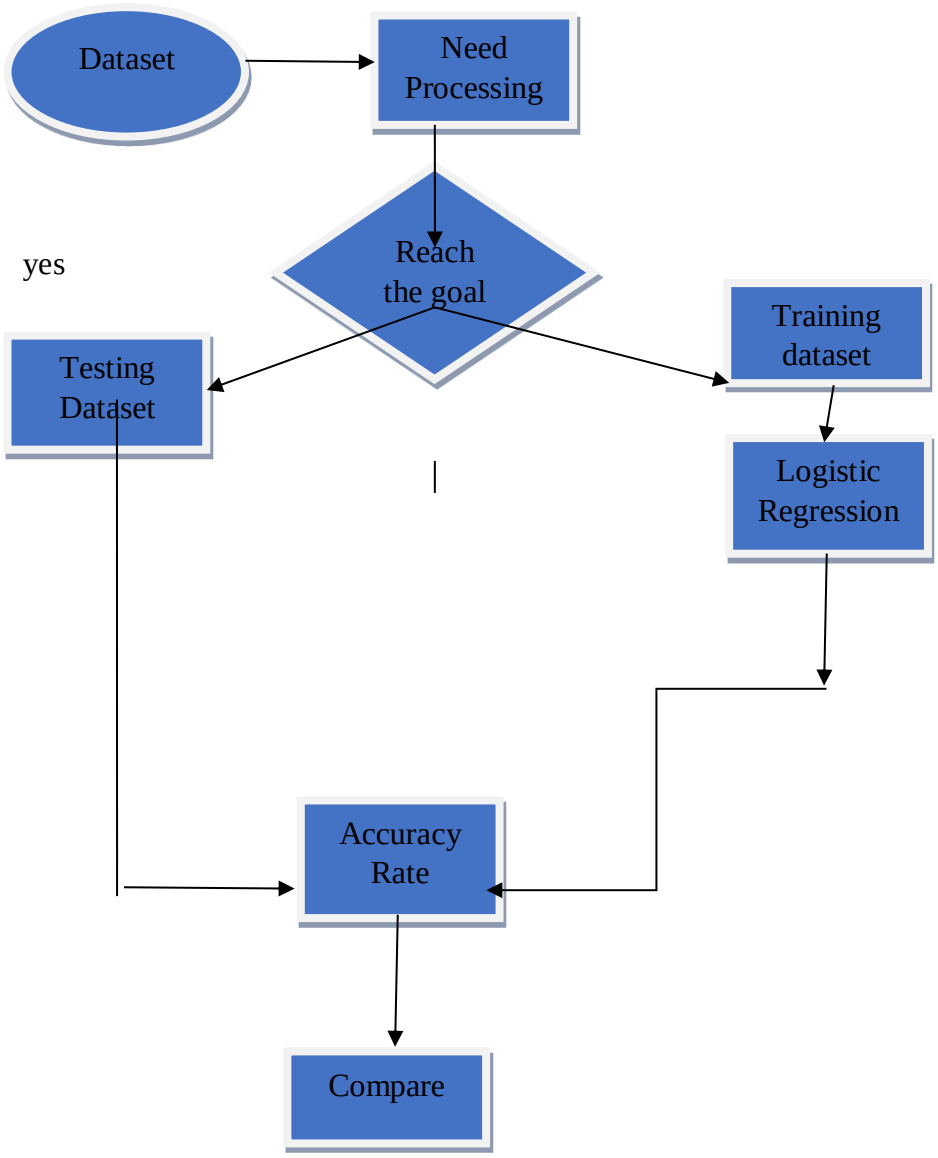


Table 2: 80 %and 20%

ATTRIBUTES	VALUE RANGE
Confusion Matrix	**225 74** **75 206**
Accuracy	**0.51**
Sensitivity	**0.28**
Specificity	**0.71**
Precision	**0.46**

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621

Table 2: 66% and 34%

ATTRIBUTES	VALUE RANGE
Confusion Matrix	123 285 215 386
Accuracy	0.50
Sensitivity	0.36
Specificity	0.57
Precision	0.30

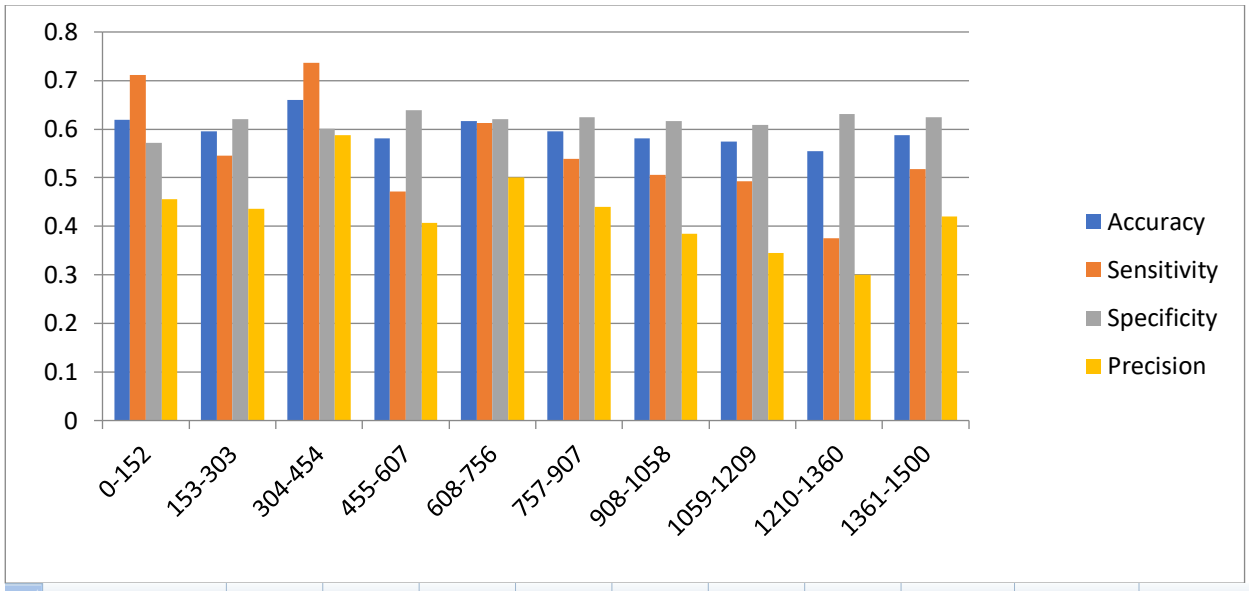
Table 2: 50 % and 50%

ATTRIBUTES	VALUE RANGE
Confusion Matrix	203 427 313 546
Accuracy	0.50
Sensitivity	0.39
Specificity	0.56
Precision	0.32

Table 4 = 10-fold cross-validation

TEST CASES	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION
01	0.61	0.71	0.57	0.45
02	0.59	0.54	0.62	0.43
03	0.66	0.73	0.60	0.58
04	0.58	0.47	0.63	0.40
05	0.61	0.61	0.62	0.5
06	0.59	0.53	0.62	0.44
07	0.58	0.50	0.61	0.38
08	0.57	0.49	0.60	0.35
09	0.55	0.37	0.63	0.30
10	0.58	0.51	0.62	0.42

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621



	A	B	C	D	E	F	G	H	I	J	K
1		0-152	153-303	304-454	455-607	608-756	757-907	908-1058	1059-1209	1210-1360	1361-1500
2	Accuracy	0.619	0.595	0.66	0.581	0.617	0.595	0.581	0.574	0.555	0.588
3	Sensitivity	0.712	0.545	0.737	0.471	0.612	0.539	0.506	0.493	0.375	0.518
4	Specificity	0.572	0.621	0.601	0.639	0.62	0.625	0.617	0.609	0.631	0.625
5	Precision	0.456	0.436	0.588	0.407	0.5	0.44	0.385	0.345	0.3	0.42

CONCLUSIONS

This paper uses multiple linear regressions (MLR) to predict health issues of student's during covid-19. The author collected the data from UCI Machine Learning Repository based on that we made a relationship between the dependent variable and the independent variable after that we perform cross-validation for more accuracy. The Author move to the Confusion matrix where we compare the actual target values with those predicted by the machine learning model, after checking the cross-validation. The Author predict the accuracy, sensitivity, and specificity for ½, ⅔ and 1/4 set of data and 10-fold cross-validation in this model.

Moreover, the best value for the number of activities in the LR algorithm was determined using an algorithm coded in Python. It was shown that when the LR method is hybridized with nine fields , the classification accuracy increases significantly.. These experiments and results were analysed carefully by author. Moreover, the author analysed the outcomes using some student's data and compared their real condition with the class that the model devoted to that students.

Future works may involve the use of other machine-learning techniques and compare their performances to the one obtained by the proposed approach.

REFERENCE

.Duduzile Ndwandwe1 and Charles S Wiysonge1,2,3. **"COVID-19 vaccines".**www.sciencedirect.com

- Ahmad, A., Garhwal, S., Ray, S. K., Kumar, G., Malebary, S. J., & Barukab, O. M. (2021). The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. *Archives of Computational Methods in Engineering*, *28*, 2645-2653.
- Sell S: How vaccines work: immune effector mechanisms and designer vaccines. Expert Rev Vaccines 2019, 18:993-1015 The study highlights the fundamentals of how vaccines work at a molecular level.

DOI:10.46647/ijetms.2023.v07i02.035 ISSN: 2581-4621

- Nagy A, Alhatlani B: An overview of current COVID-19 vaccine platforms. Comput Struct Biotechnic J 2021, 19:2508-2517 This paper applies structural biology to enable an understanding of vaccine platforms.
- Chen Y, Li L: SARS-CoV-2: virus dynamics and host response. Lancet Infect Dis 2020, 20:515-516
- Alimadadi A et al (2020) Artificial intelligence and machine learning to fight COVID-19. American Physiological Society, Bethesda Amar LA, Taha AA, Mohamed MY (2020) Prediction of the final size for covid-19 epidemic using machine learning: a case study of Egypt. Infect Dis Model 5:622–6
- Centres for Disease Control and Prevention, "Coronavirus 2019 (COVID-19) Situation Summary," Published in 2020 Available: https://www.cdc.gov/coronavirus/2019-ncov/summary. HTML background.
- S. Albahri2 &Rule A. Hamid3, Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19):A Systematic Review, https://doi.org/10.1007/s10916-020-01582-x
- Arash Heidari• Nima Jafari Navimipour, Machine learning applications for COVID-19 outbreak management, https://doi.org/10.1007/s00521-022-07424-w(0123456789
- Vahdat S et al (2021) POS-497 characteristics and outcome of COVID-19 in patients on chronic hemodialysis in a dialysiscenter specific for COVID-19. Kidney Int Rep 6(4):S215
- Zhan C, Chen J, Zhang H (2021) An investigation of testing capacity for evaluating and modeling the spread of coronavirus disease. Inf Sci 561:211–229
- Chowdhury NK et al (2021) ECOVNet: a highly effective ensemble based deep learning model for detecting COVID-19. PeerJ Comput Sci 7:e551
- Jaber MM et al (2022) Remotely monitoring COVID-19 patient health condition using metaheuristics convolute networks from IoT-based wearable device health data. Sensors 22(3):1205
- Zhang L et al (2021) Body temperature monitoring for regular COVID-19 prevention based on human daily activity recognition. Sensors 21(22):7540
- **Junaid Shuja, Eisa Alanazi2, COVID-19 open source data sets: a comprehensive survey,** https://doi.org/10.1007/s10489-020-01862-6.