Cloud Computing and Data Science
https://ojs.wiserpub.com/index.php/CCDS/

Research Article

A Study of Using Machine Learning in Predicting COVID-19 Cases

Maleerat Maliyaem, Nguyen Minh Tuan*, Demontray Lockhart, Supattra Muenthong

Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
E-mail: minh.tuan@itd.kmutnb.ac.th

Received: 16 April 2022; **Revised:** 1 July 2022; **Accepted:** 2 July 2022

Abstract: With an unprecedented challenge to combat COVID-19, the prediction of confirmed cases is very important to ensure medical aid and healthy living conditions. In order to predict confirmed cases, the current study uses a dataset prepared by the White House Office of Science and Technology Policy, which brought together companies and research to address questions concerning COVID-19. The importance of this was to identify factors that seem to affect the transmission rate of COVID-19. The focus of the current research, however, is to predict global cases of COVID-19. There have been many papers written about the prediction of confirmed cases and fatalities, but they failed to show promising results. Our research applies machine learning for predicting fatalities in the world using the COVID-19 Forecasting dataset from Kaggle. After trying several algorithms, our findings reveal that Logistic Regression, Decision Tree, KNeighbors, GaussianNB, and Random Forest algorithms provide the best predictions. Thus, the results show Random Forest as having the highest accuracy, followed by Logistic Regression and Decision Tree. The results are promising, opening up the door for further research.

***Keywords*:** COVID-19, machine learning, algorithms, Random Forest, fatality

1. Introduction

The COVID-19 disaster has become a global issue. It has adversely affected people's livelihood. Health ministries have struggled to find ways to deal with its ramifications. Medical establishments have been under pressure to deal with an alarming increase of COVID victims. To address many of the problems generated by COVID-19, researchers of various kinds have run tests and trials hoping to ascertain underlying factors that will prove useful in remedying this global threat. Some of these researchers have employed machine learning algorithms to predict the aftermath of COVID-19, using, for example, SVM and PR models to predict COVID-19's aggressive risk [1]. In this paper, we would apply machine learning such as the Linear Regression, Logistic Regression, Decision Tree, Support Vector Machines (SVM), and Random Forest to predict the confirmed cases whether they get positive or not. The exact prediction could propose a new approach for doing research in looking for an effective remedy for killing the SARS-COVID virus. On the other hand, the prediction could give the government the information for vigilance about cases of positive, places for quarantine, medicine for positive cases, and so on [2]. In the paper [2], they showed the results in prediction for positive cases with machine learning and just focus on Machine Learning Regression algorithms. In the paper [3], they predict fatalities using Linear Regression and Polynomial Regression. The results showed that Polynomial Regression has better

DOI: https://doi.org/10.37256/ccds.3220221488

results than Linear Regression. So, the paper did not investigate the full approach for prediction. In this paper, we will show the results for algorithms and comparison between them to get new full sightseeing for predicting.

The COVID-19 disaster has become a global issue. It has adversely affected people's livelihood. Health ministries have struggled to find ways to deal with its ramifications. Medical establishments have been under pressure to deal with an alarming increase of COVID victims. To address many of the problems generated by COVID-19, researchers of various kinds have run tests and trials hoping to ascertain underlying factors that will prove useful in remedying this global threat. Some of these researchers have employed machine learning algorithms to predict the aftermath of COVID-19, using, for example, SVM and PR models to predict COVID-19's aggressive risk [1].

In this paper, we would apply machine learning such as the Gaussian Naïve Bayes, KNeighbors, Logistic Regression, Decision Tree, SVM (class SVC), and Random Forest to predict the confirmed cases whether they get positive or not. The exact prediction could propose a suggestion for supplying medicine and delivering medical equipment in different regions. On the other hand, the prediction could give the government the information for vigilance about the cases of positive, places for quarantine, medicine for positive cases, and so on [2]. In the paper [2], they showed the results in prediction for positive cases with machine learning and just focus on Machine Learning Regression algorithms. In the paper [3], they predict fatalities using Linear Regression and Polynomial Regression. The results showed that Polynomial Regression has better results than Linear Regression. So, the paper did not investigate the full approach for prediction. In this paper, we will show the results for algorithms and comparison between them to get new full sightseeing for predicting.

2. Data set

In this paper, we get the data set from https://www.kaggle.com/lingyuxiong/covid19-forecasting (see Table 1). The data set has 8 attributes where two attributes Province/State and Country/Region are string, one is day format and the others are numerical. The data set consists of nearly 18,000 entries of which we performed a train-test split of a 7:3 ratio respectively. The records for the train part are 12,547 records and for test, part are 5,345 records. We would take an investigation with this data to find the negative, positive, or no correlation between the attributes (see Table 1). In this data set, we considered the information in the attributes by finding the relation in total death cases by performing the correlation in Table 1. As Table 1 shown, total confirmed cases have the positive correlation to total death cases. That means if the total confirmed cases and total deaths are directly proportional to each other.

Table 1. Results for prediction

Id	Province/State	Country/Region	Lat	Long	Date	Confirmed Cases	Fatalities
1	NaN	Afghanistan	33	65	1/22/2020	0	0
2	NaN	Afghanistan	33	65	1/23/2020	0	0
3	NaN	Afghanistan	33	65	1/24/2020	0	0
4	NaN	Afghanistan	33	65	1/25/2020	0	0
5	NaN	Afghanistan	33	65	1/26/2020	0	0
6	NaN	Afghanistan	33	65	1/27/2020	0	0

3. Literature reviews

Although COVID-19 has been intractable along with its concomitants, there has been a growth of research concerning how to address issues of predicting mortality rates, allocating medical resources, providing accurate

prognostic assessments, and managing data related to this global pandemic. Researchers have addressed these issues by employing machine learning algorithms such as Logistic Regression, Support Vector Machines, Random Forest, and Decision Trees as well as various hybrid models.

A hybrid prediction model [2] that consisted of mathematical and statistical approaches was used to predict the number of new confirmed COVID-19 cases using a dataset from WHO. The model was comprised of the Rate of Change, the Geometric mean, Standard Deviation, and calculating expected cases and boundaries. These were tested using the Mean Square Error and the correlation between the expected values and real values. Compared to the Bayesian Ridge regression model, it showed higher accuracy.

Python's Prophet's library has proven to be insightful [1] for predicting certain COVID-19-related outcomes one week in advance. These outcomes were confirming worldwide COVID-19 cases, deaths, and recoveries in India, Wuhan, South Korea, and Italy. This was also achieved by not tweaking any parameters or using additional regressors.

Similar to the aforementioned study, they [4] employed multiple machine learning algorithms (i.e., neural networks, random forests, gradient boosting trees, logistic regression, and support vector machines) on data from emergency care admission exams. The data was collected from 235 adult patients in Sao Paulo, Brazil. About 43 percent of patients received a positive diagnosis of COVID-19 from RT-PCR tests. Support vector machines algorithm performed the best (AUC: 85; Sensitivity: 0.68; Specificity: 0.85). The three most important variables were lymphocytes, leukocytes, and eosinophils in that order. They conclude by validating the importance of using routinely-collected data to make targeted decisions about the use of COVID-19 tests.

Trying to predict mortality rates, they [5] took a machine learning approach (e.g., XGBoost) to clinical data from a cohort of over 5,000 COVID-19 patients treated at a medical establishment in New York City. The predictors classified patients as either deceased or alive. The predictors were based on features such as age, minimum O_2 saturation during encounters, type of patient encounters, hydroxychloroquine use, and maximum body temperature.

In the interest of trying to inform the Indian government, they [6] studied the outbreak of COVID-19 by using machine learning models such as SEIR and Regression. They use these models on data collected from the official portal of India's government from January 2020 to May 2020. In paper [6], they used RMSLE to evaluate the models, which achieved 1.52 for the SEIR model and 1.75 for the regression model. They state that their research will help the government and doctors to plan more appropriately for the future.

Using multipurpose machine learning algorithms (i.e., artificial neural networks, extra trees, random forests, catboost, and extreme gradient boosting), they [3] were able to predict the risk of COVID-19 patients developing critical conditions. They used a 7:3 train-test split of a dataset including 1,040 patients with a positive RT-PCR diagnosis for COVID-19. The algorithms were trained by combining outcomes to predict the other which proved to show high predictive performance (average AUROC of 0.92, sensitivity of 0.92, and specificity of 0.82). For the multipurpose algorithms, three important variables that were identified were lymphocyte per C-reactive protein, C-reactive protein, and Braden Scale. They [3] concluded that using machine learning algorithms on routinely-collected data can be used to predict unspecific negative COVID-19 outcomes. Using hybrid machine learning methods of an Adaptive Network-based Fuzzy Inference System (ANFIS) and Multi-Layered Perceptron-Imperialist Competitive Algorithm (MLPICA), they [4] were able to predict time series of infected individuals and mortality rate with regards to the COVID-19 outbreak in Hungary. The models predicted that the outbreak and mortality rate would drop drastically by late May. This was based on sampling data split into odd and even days for which the validation was performed for 9 days, showing promising results. The MLP-ICA performed better than the ANFIS providing accurate results on validation samples.

In this paper, we apply the algorithms such as Logistic Regression, KNneighbors, Gaussian Naïve Bayes, Random Forest, and Decision Tree and we compare them to get the best result in prediction. Logistic Regression [5] is an algorithm for classification for discrete values using Sigmoid function. Logistic Regression is applied to analyze and predict the results of prior investigated variables. Logistic Regression solves the optimization problem with optional classes l_1 and l_2 or Elastic-Net regularization. Binary class l_2 penalized logistic regression minimizes the following cost function (1). Class l_1 regularized logistic regression evaluates the following optimization problem (2).

$$min_{w,e} \frac{1}{2} w^T w + C \sum_{i=1}^{n} log\left(exp\left(-y_i\left(X_i^T w + c\right)\right) + 1\right) \tag{1}$$

$$min_{w,e} \| u \|_1 + C\sum_{i=1}^{n} log\left(exp\left(-y_i\left(X_i^T w + c\right)\right)+1\right) \tag{2}$$

Random Forest is a kind of supervised learning algorithm. It is a useful and flexible algorithm because it is implicit and diverse, applied for both classification and regression tasks. Random Forest adds by itself additional randomness to the model and searches for the best features while growing the tree. Random Forest is a meta estimator that fits a large number of sizes of data and estimates the predictive accuracy.

Decision Tree is also a kind of supervised learning model and could be applied in two kinds of classification and regression problems. Building a Decision Tree based on training prior dataset is confirmation their requirements and order. The Decision Tree could work with features formed from categorical or numeric attributes.

Gaussian Naïve Bayes is usually applied in data with continuous parameters. The class i and input values c_i, x_i belong to standard distribution with expecting μ_{ci} and variance σ_{ci}^2 as form (3)

$$p\left(x_i \mid c\right) = p\left(x_i \mid \mu_{ci}, \sigma_{ci}^2\right) = \frac{1}{\sqrt{2\pi\sigma_{ci}^2}} exp\left(-\frac{\left(x_i - \mu_{ci}\right)^2}{2\sigma_{ci}^2}\right) \tag{3}$$

KNeighbors is applied in supervised and unsupervised methods to predict the labels from a training set. This method is based on k different nearest neighbors to classify the data by finding nearest observation in classifying or regression problems. KNeighbors could be used as very populous in some classifying two or more than two classes. KNeighbors used approximation methods to calculate the classes in the local region and expand the target point to neighbor target to get the full labels.

The Support Vector Machines, a kind of supervised learning method, is usually applied in classification and detection problems. It is useful in the data with binary or multi-class targets. There are many classes in the structure such as SVC, NuSVC, and LinearSVC, but they are all capable to perform very perfectly in solving the data with the above two tasks.

4. Experiments

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17892 entries, 0 to 17891
Data columns (total 8 columns) :

#	column	Non-Null Count	Dtype
0	Id	17892 non-null	int64
1	Province/State	8190 non-null	object
2	Country/Region	17892 non-null	object
3	Lat	17892 non-null	float64
4	Long	17892 non-null	float64
5	Date	17892 non-null	object
6	Confirmed Cases	17892 non-null	int64
7	Fatalities	17892 non-null	int64

dtypes: float64 (2), int64 (3), object (3)
memory usage: 1.1+ MB
None

Figure 1. Information for data

The data set consists of 12,212 instances with 5 contributions about forecast ID, Province, Country, Latitude (lat), Longtitude (long), Date, and Region. we divided it into 2 parts concluding with the train set and test set. In the future,

we choose a bigger set and apply a random forest, a decision tree. The data (Figure 1) has many kinds of unformed attributes. With the column day, we change to an integer by omitting the dash between the number and applying cast integer. In this paper, we divided the data into two types of attributes. With the string and numerical attributes, we apply OnehotLabel to change to numeric and applied Standard-Scalar to process data before establishing Features. After that, we put all the attributes into Features for predicting. With the attribute Fatalities, we turn into a Label integer for prediction. We built models with the above algorithms and choose the best accuracy for predictions such as Logistic Regression, Gaussian Naïve Bayes, KNeighbors, Random Forest, SVM, and Decision Tree. The steps for prediction are also shown in Figure 2. To predict the data, we apply machine learning with many models in machine learning and choose five high accuracy models for prediction. In this paper, we apply machine learning with the algorithms for five models Logistic Regression, Random Forest, Gaussian Naïve Bayes, SVM, KNeighbors, and Decision Tree. The results have been shown in Table 2. In this table, the Random Forest has the best performance. To evaluate the accuracy of the model, we apply the formula (4-7).

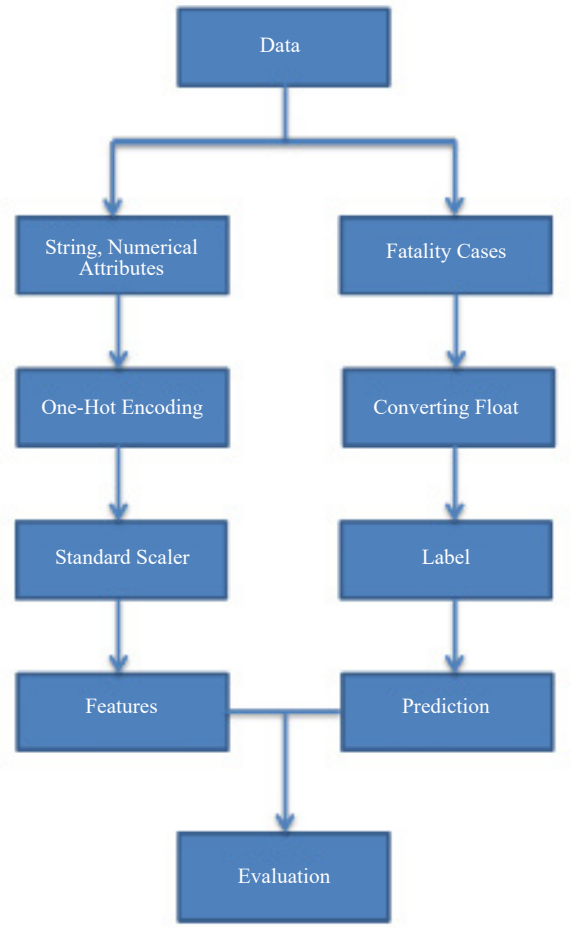

Figure 2. Steps for analyzing data

5. Results

$$Accuracy = \frac{\sum_{i=1}^{n} T_{iV}}{\sum_{i=1}^{n} T_{iV} + \sum_{j=1}^{m} F_{jV}} \quad (4)$$

Where n, m are numbers of classes, T_{iV} is a true value of prediction at class i; F_{jV} is a false value of label at class j.

$$Precision = \frac{\sum_{i=1}^{n} T_{iP}}{\sum_{i=1}^{n} T_{iP} + \sum_{j=1}^{m} F_{jP}} \tag{5}$$

Where T_{iP} is the true positive at class i, and F_{iP} is false positive at class i. F_{jN} is false negative at class j.

$$Recall = \frac{\sum_{i=1}^{n} T_{iP}}{\sum_{i=1}^{n} T_{iP} + \sum_{j=1}^{m} F_{jN}} \tag{6}$$

$$F_1 = 2 \times \frac{Precision?\ Recall}{Precision + Recall} \tag{7}$$

Table 2. Results for prediction

Models	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.895	0.039	0.032	0.034
KNeighbors	0.895	0.064	0.038	0.045
GaussianNB	0.509	0.022	0.064	0.025
Decision Tree	0.939	0.090	0.074	0.078
Random Forest	0.950	0.086	0.100	0.087
SVM	0.901	0.058	0.060	0.055

Table 3. Some example results for prediction

Label	Random Forest	Decision Tree	Logistic Regression
8	5	3	4
167	98	7	1
2	0	0	121
149	152	184	1
148	172	184	1

To get a clear performance, we perform the *Precision*, *Recall*, and *F1-score* shown in Table 2. With the development of epidemic diseases, COVID-19 cases are more and more increasing day by day, we hope to apply machine learning could predict the fatalities is also one of the ways to forbid the situation [6]. We would like to strengthen machine learning in prediction with a suitable method. Anyways, Random Forest established the function that is discrete values as Logistic Regression. Based on Table 3, we could also see the limit of GaussianNB and KNeighbors in predicting the values for the cases with this data. The COVID-19 cases sometimes could predict the harmfulness by

checking the Xrays images of the patients [7]. To show more the performance of the prediction, we count the values and express the prediction by diagram from Table 3. In the performance of the algorithms, we could show the values of prediction and estimate the predictions by counting the exact values we could apply the prediction to every country and for every case in the data. After training the models, we showed the number of predictions. For example, the results of Logistic Regression, Random Forest, and Decision Tree prediction have shown in Table 3. Figure 3, using a scatter plot, showed them the full prediction using Random Forest and the labels are almost covered by predicting values. Figure 4 is another performance in predicting using the Decision Tree model.

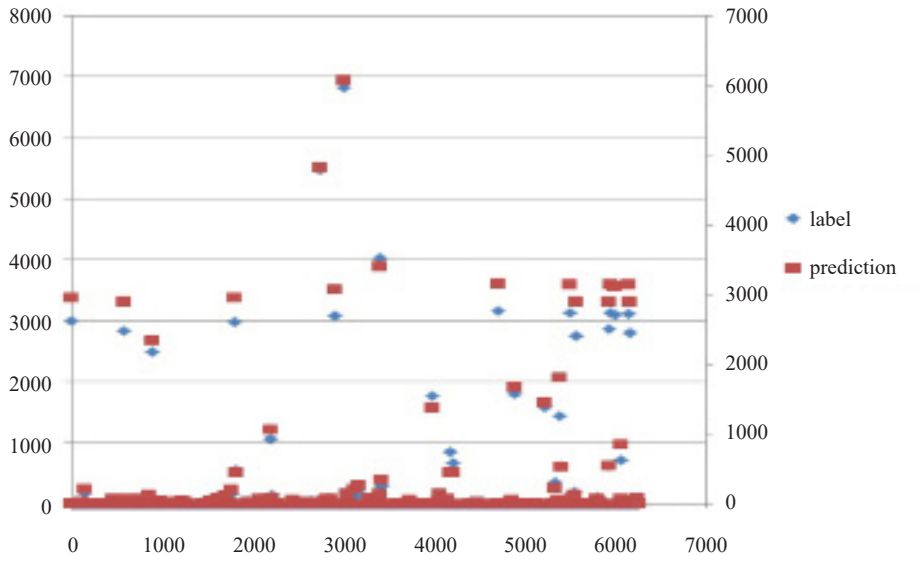

Figure 3. Prediction using Random Forest

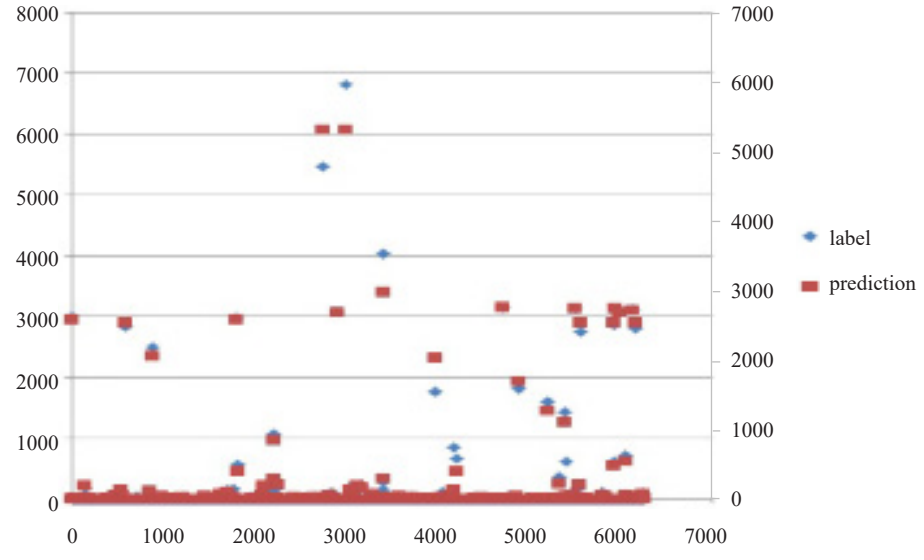

Figure 4. Prediction using Decision Tree

6. Conclusion and future work

The situation of COVID-19 suddenly came has made a broadly affected to people's health and medical supply for COVID-19 victims. In this paper, we have proposed six models for predicting COVID-19 [8]. Random Forest has shown the best result for analyzing the data with the highest accuracy in predicting COVID-19 fatalities. Machine learning has shown the different methods for guessing the output. Estimation could help the governments to establish the facilities for solving problems relating to the epidemic patients. Pandemics rising could harm the patients' health in general and could kill people to bring back the pangs of parting for left humans. A good prediction could also help many sanitation organizations and other offices to propose a strategy or stratagem for facing pandemics. The paper showed the prediction using machine learning to analyze and guess for the fatality cases [9]. In paper [10], they supply the remedy for proposing methods to cure COVID-19 cases. In this paper, we have made a prediction i n COVID cases in different regions and then process the needs for curing patients. In paper [11], they connected the data extracted from the mobile application that could contain HAR-Image for analyzing the features and summarized that HAR-Image is the effective feature for activity recognition. In this paper, we collected a data set and applied deep machine learning to guess the region with COVID cases and deliver the suitable medicine utilities. This is also the basic research for the new researchers to continue with these models for prediction. This paper is with aims to help the medical center and nutrition center to find the new vaccine and new curing methods for COVID-19 patients.

Conflict of interest

The authors hereby state that there is no conflict of interest for this article.

References

[1] Sirage ZA. Analysis and forcasting the outbreak of COVID-19 in Ethiopia using machine learning. *European Journal of Computer Science and Information Technology*. 2020; 8(4): 1-13.

[2] Tamer Sh.M. A novel machine learning based model for COVID-19 prediction. *International Journal of Avanced Computer Science and Applications*. 2020; 11(11): 523-531.

[3] Manpinder S, Saiba D. Prediction of number of fatalities due to COVID-19 using Machine Learning. *2020 17th India Council International Conference*. New Delhi, India: IEEE; 2020.

[4] Roseline OO, Joseph BA. *Machine Learning Prediction For COVID-19 Pandemic In India*. MedRxiv [Preprint]. 2020. Available from: doi: 10.1101/2020.05.20.20107847.

[5] Sudhir B, Ajit SS, Amit T, Bhoopendra P, Jyot-sna S, Sanjay S, et al. Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters. *Ibnosina Journal of Medicine and Biomedical Sciences*. 2020; 12(2): 123.

[6] Zhao HW, Naveed NM, Alyssa M, Tiffany AR, Cote MJ, Rebecca SB, et al. COVID-19: Short term prediction model using daily incidence data. *PLoS ONE*. 2021;16(4): e0250110. Available from: doi: 10.1371/journal.pone.0250110.

[7] Haritha D, Swaroop N, Mounika M. Prediction of COVID-19 Cases Using CNN with Xrays. *2020 5th International Conference on Computing, Communication and Security (ICCCS)*. Patna, India: IEEE; 2020.

[8] Vartika B, Anand SJ, Pooja P. A comparative study of Machine Learning Models for COVID-19 prediction in India. *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*. Chennai, India: IEEE; 2020.

[9] Soufiane H, Oussama EL G, Bouchaib C, Hassan O, Abdelhadi R. Optimization of Machine Learning Algorithms Hyper-Parameters for Improving the Prediction of Patients Infected with COVID-19. *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*. Kenitra, Morocco: IEEE; 2020.

[10] Gianni D'A, Francesco P. Discovering genomic patterns in SARS-CoV-2 variants. *International Journal of Intelligent systems*. 2020; 35(11): 1680-1698.

[11] Gianni D'A, Francesco P. Enhancing COVID-19 tracking apps with human activity recognition using a deep convolutional neural network and HAR-images. *Neural Computing and Applications* [Preprint] 2021. Available from: doi: 10.1007/s00521-021-05913-y.