MANAS Journal of Engineering
ISSN 1694-7398 | e-ISSN 1694-7398

Volume 8 (Issue 2) (2020) Pages 144-150

Forecasting COVID-19 Cases based on mobility

Mehmet Şahin

Department of Industrial Engineering, Iskenderun Technical University, 31200 Iskenderun, Turkey,
mehmet.sahin@simon.rochester.edu, ORCID: 0000-0001-7078-7396

ABSTRACT

Countries struggling to overcome the profound and devastating effects of COVID-19 have started taking steps to return to the "new normal." Any accurate forecasting can help countries and decision-makers make plans and decisions in returning to normal life. In this regard, it is needless to mention the criticality and importance of accurate forecasting. In this study, daily cases of COVID-19 are estimated based on mobility data, considering the proven human-to-human transmission factor. The data of seven countries, namely Brazil, France, Germany, Italy, Spain, the United Kingdom (UK), and the United States of America (USA), are used to train and test the models. These countries represent around 57% of the total cases in the whole world. In this context, various machine learning algorithms are implemented to obtain accurate predictions. Unlike most studies, the predicted case numbers are evaluated against the actual values to reveal the methods' real performance and determine the most effective methods. The results indicated that it is unlikely to propose the same algorithm for forecasting COVID-19 cases for all countries. Also, mobility data can be enough the predict the COVID-19 cases in the USA.

ARTICLE INFO

Research article

Received: 15.07.2020
Accepted: 13.12.2020

Keywords:
COVID-19;
forecasting;
mobility;
human-to-human transmission;
coronavirus

1. Introduction

Coronavirus disease 2019 (COVID-19) pandemic, causing 6,799,713 confirmed cases and 397,388 deaths in 216 countries as of June 7, 2020, has had impacts that are difficult to recover in the short term [1]. The outbreak affected people's health and lifestyles and brought problems such as the global economic downturn, psychological distress, and adverse effects on daily activities [2]. To minimize or eliminate such problems, countries have taken various public policy decisions regarding health, social, and economical. When making such short- or long-term decisions, predictions regarding the outbreak are crucial inputs.

Considering valid parameters in estimates significantly affects the accuracy of the results, the mobility variable is considered significant in predicting COVID-19 cases. This situation can be explained as follows. Various factors contribute to the global spread of infectious diseases, including increased speed and reach of human mobility and increased volumes of trade and tourism [3]. As an infectious disease, COVID-19 caused the suspension of international and domestic flights and lockdown restrictions in many countries [4, 5]. Such radical decisions, proving that mobility is a significant parameter, were made to minimize mobility and hence human-to-human transmission [6]. Human-to-human transmission happens through the respiratory droplets released when infected patients sneeze or cough [7].

Recent studies also claimed that mobility and human-to-human transmission are effective factors in the spread of COVID-19 [8-10]. In addition, Kraemer, Yang [11] analyzed the impact of human mobility in China and found that the mobility statistics provided a precise record of the spread of COVID-19 among Chinese cities in early 2020. Tagliazucchi, Balenzuela [12] considered mobility through utilizing cell phone data to model the expansion of the COVID-19 in Argentina.

It can be concluded that the mobility of people and the population are the primary sources of human-to-human transmission. Therefore, these two factors are included in the prediction models in this study. This study's main objective is to assess the impact of mobility and determine the most effective methods in predicting daily COVID-19 cases. In this context, numerous machine learning models are formed. To train and test the models, the data of seven countries, namely Brazil, France, Germany, Italy, Spain, the United Kingdom (UK), and the United States of America (USA), are used. These countries had approximately 57% of the total number of cases in the world as of May 26, 2020. Thus, more than half of the total COVID-19 cases are involved in the analysis.

Although Russia has many confirmed cases, it was excluded from the analysis as the detailed mobility data do not exist for the country.

There have been some studies based on forecasting COVID-19 cases. Yousaf, Zahir [13] estimated COVID-19 cases, recoveries, and deaths in Pakistan through Auto-Regressive Integrated Moving Average Model (ARIMA). In addition, some other studies predicted daily cases and deaths, alone or together in different countries. Salgotra, Gandomi [14] predicted confirmed cases and death cases in India; Ayinde, Lukman [15] proposed and compared several models for COVID-19 cases in Nigeria; Mandal, Jana [16] predicted cases in three states on India; Parbat and Chakraborty [17] implemented support vector regression to predict cases in India; Chimmula and Zhang [18] predicted COVID-19 transmission in Canada; and Fanelli and Piazza [19] predicted cases in China, France, and Italy. It can be inferred that different models were adopted to forecast COVID-19 cases in one country in general. Unlike other studies, in the present study, seven countries are included in the analysis, and the predicted cases are evaluated against actual cases. Also, updated real data are used. Considering the data size is significant in analyzing machine learning methods, the present study also contributes to the literature using the most recent and considering a broader timeframe than earlier studies. To the best of my knowledge, no reports or studies are available in literature at the time of preparing this study, which uses the most up-to-date and such broad data, addresses several countries, and evaluates the prediction results against real cases. Therefore, the purpose of this study constitutes the first successful and complete in determining the most effective methods for predicting daily COVID-19 cases considering seven countries by using updated data and evaluating the predicted results with real data.

2. Materıals and methods

2.1. Study Area

Seven countries, namely Brazil, France, Germany, Italy, Spain, the UK, and the USA, are considered in the analysis. As mentioned earlier, these countries represent approximately 57% of the total cases worldwide as of preparing the present study. These are the countries most affected by the COVID-19 pandemic in terms of the total case. Figure 1 presents the daily cases of each country. Most cases were located in the USA, followed by Spain, Brazil, the UK, Italy, Germany, and France as of May 20, 2020 [20].

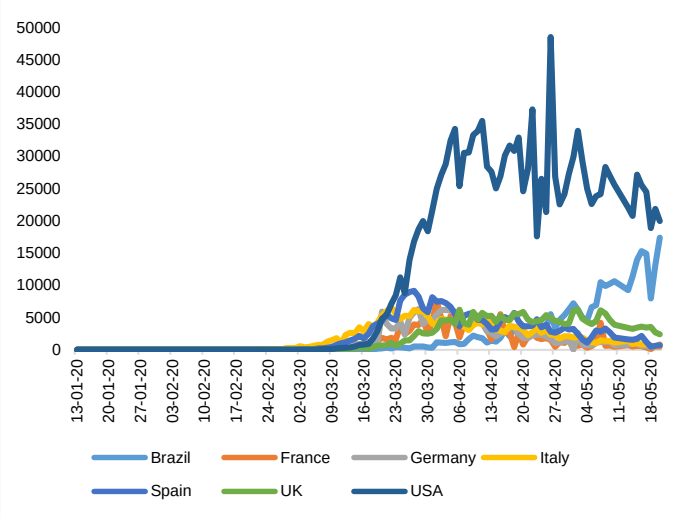

Figure 1. *Daily COVID-19 cases in each country*

2.2. Data

The mobility data were obtained from Apple Mobility Trends Reports (https://www.apple.com/covid19/mobility). Figures 2, 3, and 4 demonstrate the comparison of driving, transit, and walking indices of countries, respectively. The data for May 11 – 12 is not available. Therefore, the data from Jan 13, 2020 to May 10, 2020 were used for training the models, and the data from May 13, 2020 to May 20, 2020 were used to test the models. The beginning of the training data was determined, considering the incubation period of COVID-19 changes from 1 day to 14 days [21]. Besides, the daily cases were obtained from the EU Open Data Portal [20].

Figure 2 illustrates that countries' driving index differs; however, there is a similar trend for all. In other words, a sharp decrease in March and an increasing tendency afterward can be observed in the figure. Remaining the highest for a while and falling to the lowest values, Spain's index can be the summary of the story of this country once being the epicenter of COVID-19.

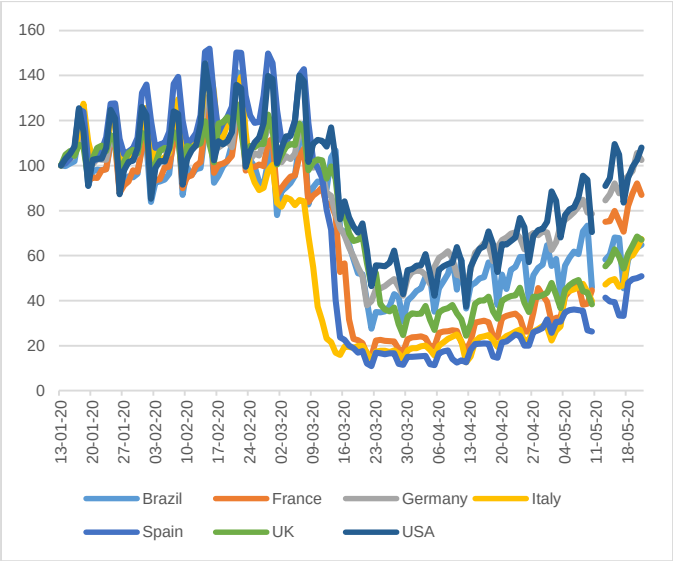

Figure 2. *Driving index of countries*

Figure 3 presents the comparison of transit indices of the countries. Higher values for Germany and Spain in the early stages are seen. Besides, the highest transit indices were measured in Germany in the later period.

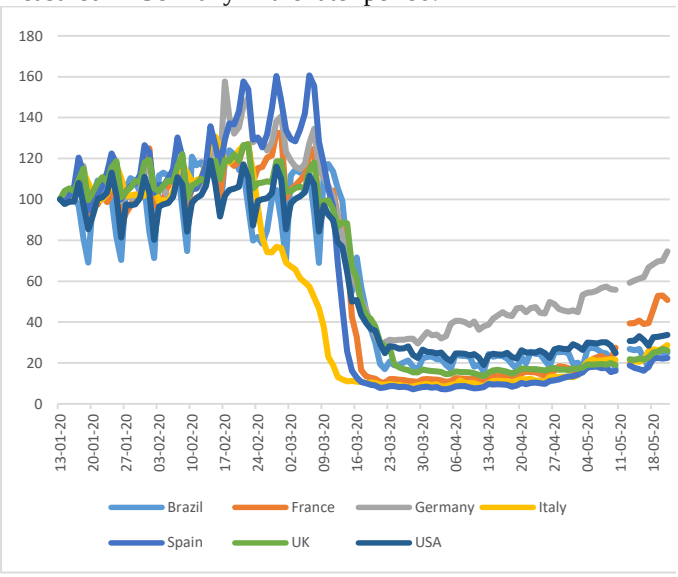

Figure 3. *Transit index of countries*

Figure 4 illustrates the comparison of the walking indices of the countries. This trend is similar to the previous two. The highest values in the UK in the early period, in Spain in the following period and Germany in the near future stand out.

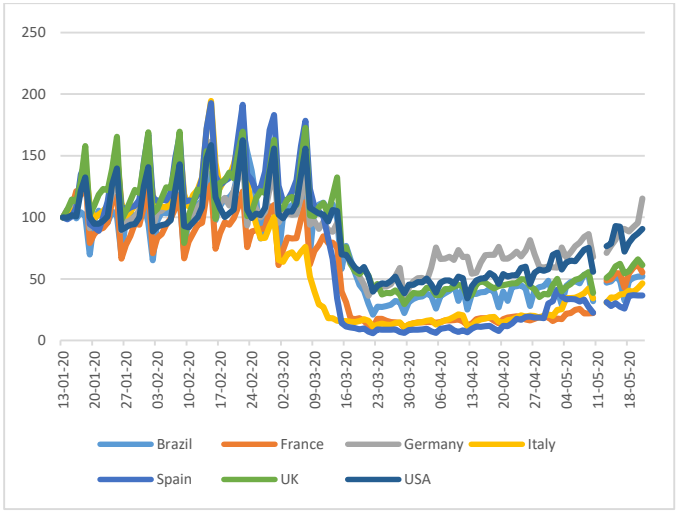

Figure 4. *Walking index of countries*

2.3. Description of the Methods

The methods are briefly described as follows:

1. *Linear Regression Models* deliver an output (prediction) based on a linear combination of inputs. The significant advantages of these models are ease of application and interpretability. However, these models can deliver misleading predictions if the relationship between input and output cannot be reasonably interpreted through a linear function.

2. *Gaussian Process Regression Models* are based on a practical and probabilistic approach to learning [22]. Gaussian process regression (GPR) models represent a probabilistic method appropriate for nonlinear regression problems. These models can deal with complex problems that have different features, including small samples, nonlinearity, and high dimensions. They can detect uncertainty and deliver compelling predictions.

3. *Support Vector Machine (SVM)* is implemented for regression, time-series, and classification problems owing to its optimal global capacity, flexibility, forecasting ability, and minimal overfitting issues being one of the common issues in modeling high-dimensional data [23].

4. *Decision Tree* defines a dependent variable as a function of numerous independent variables. As practical algorithms, Decision Trees allow dealing with different response data types, including numeric, categorical ratings. Also, they can process missing data in both independent and dependent variables. Boosted and bagged procedures are the two most commonly used approaches. Boosted is for joining multiple classifiers to deliver better

performance compared to the individual classifier alone. Bagged is for producing several versions of a predictor through bootstrap replicates of the training data set and integrating them to better accuracy. These two procedures differ in how data are resampled [24].

5. *Ensemble Learning (EL)* comprises various learners that help to deliver accurate predictions for a given problem. Less prone to a potential data overfitting problem than a single learner and improved generalization abilities make it preferable [25].

2.4. Implementation of the Algorithms

The mobility indices and populations of the countries are used as inputs to reveal whether these are enough to forecast daily COVID-19 cases through machine learning techniques. In this context, numerous machine learning methods are implemented. The training data are used to train each approach. Then, each model's performance is evaluated against actual data in each country observed from May 13, 2020 to May 20, 2020. Thus, the models are evaluated in general and country-specific. The performance of each model is assessed based on the mean absolute percentage error.

3. Results and discussion

The results showed that it is impossible to forecast daily COVID-19 cases in all countries based on the mobility index through machine learning methods. This outcome may result from the spread of COVID-19 cases to be chaotic or an indication that the mobility index and population are not sufficient as inputs. The first case seems more relevant. Arias Velásquez and Mejía Lara [26] claimed that the COVID-19 pandemic has behaved in changing ways and unpredictable in all countries, depending on the factors used in its treatment. It is a fact that every country has written its own story during the COVID-19 pandemic. Countries differ in factors such as treatment methods, monitoring the number of infected patients, adequate health infrastructure, mortality rates, and the number of tests performed. All these factors directly or indirectly determine the number of COVID-19 cases in a country. Therefore, it can be inferred that it is more reasonable to evaluate each country separately with the results obtained.

The most accurate forecasting results were obtained for the USA. Table 1 presents the actual and predicted COVID-19 cases from May 13, 2020 to May 20, 2020 in this country.

Table 1. *Confirmed COVID-19 cases and forecasted cases in the USA*

Date	Confirmed Cases	Rational Quadratic GPR	Linear Regression	Fine Tree	Cubic SVM	Ensemble Boosted Trees
5/13/2020	22048	23896	29392	24781	33100	24703
5/14/2020	20782	23519	30250	24781	33008	24703
5/15/2020	27143	22284	34311	24781	34630	23880
5/16/2020	25508	21777	34939	24781	38595	24703
5/17/2020	24487	23781	29449	28857	36104	24703
5/18/2020	18873	19360	30271	24781	33083	24703
5/19/2020	21841	17847	32074	24781	34357	24703
5/20/2020	19970	16229	33054	24781	34919	23880

To analyze the errors, mean absolute percentage errors were calculated, as given in Table 2. To be noted that only the best performing approach for each method group (linear regression, Gaussian process regression, support vector machine, decision tree, and ensemble learning) is given. The results reveal that the rational quadratic GPR provided the most accurate predictions, followed by ensemble boosted trees and fine trees, considering that low mean absolute percentage error values indicate better estimations. Also, the percentage error of Cubic SVM is higher than 50%, meaning that forecasting is inaccurate.

Table 2. *The mean absolute percentage errors of predictions in the USA*

Rational Quadratic GPR	Linear Regression	Fine Tree	Cubic SVM	Ensemble Boosted Trees
12.07%	41.91%	16.24%	55.34%	13.82%

In addition, the lowest mean absolute percentage errors were provided for Brazil, the United Kingdom, and Germany by linear regression (20.69%), linear regression (30.14%), and rational quadratic GPR (47.03%), respectively. The errors for Spain, Italy, and France were at unacceptable levels. In other words, the machine learning methods based on the mobility index data could not provide accurate results for these three countries. The forecasting accuracy might be improved by including more parameters in the analysis.

Some results can be outlined. First, numerous machine learning methods were implemented to forecast daily COVID-19 cases in seven countries. Based on these countries' results, it is not possible to recommend the same model for all countries. In other words, no model provided accurate predictions for all countries. This outcome is consistent with the study conducted by [27]. Shahid, Zameer [27] used machine learning methods to forecast COVID-19 cases and evaluated different models' performance, including deep learning. In this regard, they considered ten countries, namely Brazil, China, Germany, India, Israel, Italy, Russia, Spain, the UK, and the USA. According to their results, long short-term memory (LSTM) provided the best predictions for China's confirmed cases and deaths. However, primarily, the present study results reveal that rational quadratic GPR can be used to predict daily COVID-19 cases in the USA. This result may be consistent with the reason for the studies considering just one country in their analyses. In this context, Tomar and Gupta [28] implemented methods such as LSTM to predict the number of COVID-19 cases in India and posited that the proposed method was effective. Arora, Kumar [29] applied deep learning methods such as recurrent neural network-based LSTM to forecast and analyze COVID-19 positive cases in India, and the proposed method produced high accuracy with an error of less than 3% for daily prediction results and less than 8% for weekly results. Alzahrani, Aljamaan [30] implemented the ARIMA model to predict the spread of the COVID-19 pandemic in Saudi Arabia and found that the proposed ARIMA model outperformed other models. The results also indicated that mobility data were not enough the predict cases in Spain, Italy, and France. Several other studies implemented machine learning approaches for different countries. Ribeiro, da Silva [31] implemented machine learning methods for estimating the cumulative confirmed COVID-19 cases in Brazil and claimed that support vector regression and stacking-ensemble learning model could be used to estimate COVID-19 cases. Sujath, Chatterjee [32] applied machine learning methods to forecast cases in India and posited that multilayer perceptron provided more accurate results than linear regression. Tuli, Tuli [33] implemented machine learning models to forecast the growth and trend of COVID-19 pandemic considering various countries and obtained different results for each country. Wang, Zheng [34] applied logistic model and machine learning methods to predict the COVID-19 trend globally, in Brazil, India, Indonesia, Peru, and Russia and estimated that the global peak would be in late October. The number of studies can be extended. Last, machine learning methods can handle extensive data and provide successful results. A larger dataset might improve the accuracy of the forecasting.

The present study might have some limitations. First, the mobility parameters that reflect the impact of human-to-human transmission were considered in this study. Although the predicted results indicated that the mobility variables were enough to estimate COVID-19 cases in the USA, more variables might improve forecasting results' accuracy. Meteorological parameters were found to be effective in COVID-19 cases [21, 35-38]. However, the fact that the COVID-19 pandemic has affected countries without regard to their development level reflects that the use of many parameters, such as economical in the COVID-19 prediction, can be misleading and unsuccessful. Second, each country has followed different strategies, and its strategies determine the expansion of COVID-19 cases. Thus, not including the specific actions taken by countries might cause inaccurate predictions in some countries. Last, due to machine learning methods producing significant results based on extensive data, conducting similar analysis after a while may increase the accuracy of results. Overall, it can be concluded that it may not be easy to propose a general model for forecasting COVID-19 cases in each country at the moment.

To sum up, the USA's recent protests have proven the mobility effect on the spread of COVID-19. Standing in the crowd for a long time increases the risk of transmission of COVID-19. It is believed that the protests created chaos, violence, and protests that can further trigger cases of COVID-19 [39]. The present study and tangible cases prove that the mobility data are crucial inputs on forecasting COVID-19 cases. Admittedly, additional variables may be required for higher accurate predictions.

4. Conclusions

Unlike other studies, this study analyzed whether COVID-19 cases could be predicted based on mobility data alone. In this context, various machine learning models were implemented to forecast daily COVID-19 cases in Brazil, France, Germany, Italy, Spain, the UK, and the USA. Using the real data of these countries, the models were trained and tested with the actual number of cases.

The results indicated that it was not possible to use the same model for all countries. This outcome is in line with the real-life situation as each country has followed different strategies to fight the COVID-19 pandemic. The results also revealed that the rational quadratic GPR could be used to predict COVID-19 cases in the USA. It is worth reminding that this model has achieved this only with mobility and population data. However, the mobility and population data alone could not yield useful results in estimating the number of cases in Spain, France, and Italy.

Including some other parameters such as meteorological and economical in the models, using different models, and analyzing each country separately may increase the results' effectiveness. These may be situations to consider for future work.

References

[1]. World Health Organization, Coronavirus disease (COVID-19) Situation Report–139, (2020).

[2]. El Zowalaty M.E., Järhult J.D., "From SARS to COVID-19: A previously unknown SARS-CoV-2 virus of pandemic potential infecting humans–Call for a One Health approach", One Health, (2020), 100124.

[3]. Findlater A., Bogoch I.I., "Human mobility and the global spread of infectious diseases: a focus on air travel", Trends in parasitology, 34(9), (2018), 772-783.

[4]. Alandijany T.A., Faizo A.A., Azhar E.I., "Coronavirus disease of 2019 (COVID-19) in the Gulf Cooperation Council (GCC) countries: Current status and management practices", Journal of Infection and Public Health, 13(6), (2020), 839-842.

[5]. Al-Tawfiq J.A., Memish Z.A., "COVID-19 in the Eastern Mediterranean Region and Saudi Arabia: prevention and therapeutic strategies", International Journal of Antimicrobial Agents, 55(5), (2020), 105968.

[6]. Zhu Y., et al., "The mediating effect of air quality on the association between human mobility and COVID-19 infection in China", Environmental Research, 189, (2020), 109911.

[7]. Wang J., Du G., "COVID-19 may transmit through aerosol". Irish Journal of Medical Science (1971 -),. 189(4), (2020), 1143-1144.

[8]. Phan, L.T., et al., "Importation and human-to-human transmission of a novel coronavirus in Vietnam", New England Journal of Medicine, 382(9), (2020), 872-874.

[9]. Riou J., Althaus C.L., "Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020", Eurosurveillance, 25(4), (2020), 2000058.

[10]. Muhammad S., Long X., Salman M., "COVID-19 pandemic and environmental pollution: A blessing in disguise?", Science of The Total Environment, 728, (2020), 138820.

[11]. Kraemer M.U.G., et al., "The effect of human mobility and control measures on the COVID-19 epidemic in China". Science, 368(6490), (2020), 493.

[12]. Tagliazucchi E., et al., "Lessons from being challenged by COVID-19", Chaos, Solitons & Fractals, 137, (2020), 109923.

[13]. Yousaf M., et al., "Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan", Chaos, Solitons & Fractals, 138, (2020), 109926.

[14]. Salgotra R., Gandomi M., Gandomi A.H., "Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming". Chaos, Solitons & Fractals, 138, (2020), 109945.

[15]. Ayinde K., et al., "Modeling Nigerian Covid-19 cases: A comparative analysis of models and estimators", Chaos, Solitons & Fractals, 138, (2020), 109911.

[16]. Mandal M., et al., "A model based study on the dynamics of COVID-19: Prediction and control", Chaos, Solitons & Fractals, 136, (2020), 109889.

[17]. Parbat D., Chakraborty M., "A python based support vector regression model for prediction of COVID19 cases in India". Chaos, Solitons & Fractals, 138, (2020), 109942.

[18]. Chimmula V.K.R. Zhang L., "Time series forecasting of COVID-19 transmission in Canada using LSTM networks", Chaos, Solitons & Fractals, 135, (2020), 109864.

[19]. Fanelli D., Piazza F., "Analysis and forecast of COVID-19 spreading in China, Italy and France", Chaos, Solitons & Fractals, 134, (2020), 109761.

[20]. EU Open Data Portal. COVID-19 Coronavirus data 2020 [cited 2020 5/29/2020]; Available from: https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-geographic-disbtribution-worldwide.xlsx.

[21]. Şahin M., "Impact of weather on COVID-19 pandemic in Turkey", Science of The Total Environment, 728, (2020), 138810.

[22]. Williams C.K.I., Rasmussen C.E., "Gaussian processes for machine learning", MIT press Cambridge, MA., 2, (2006).

[23]. Thissen U., et al., "Using support vector machines for time series prediction", Chemometrics and Intelligent Laboratory Systems,. 69(1), (2003), 35-49.

[24]. Moisen G.G., "Classification and Regression Trees, in Encyclopedia of Ecology, S.E. Jørgensen and B.D.

Fath, Editors, Academic Press: Oxford, (2008), 582-588.

[25]. Wu X., et al., "Top 10 algorithms in data mining", Knowledge and Information Systems, 14(1), (2008), 1-37.

[26]. Arias Velásquez R.M. Mejía Lara J.V., "Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression", Chaos, Solitons & Fractals, 136, (2020), 109924.

[27]. Shahid F., Zameer A., Muneeb M., "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM", Chaos, Solitons & Fractals, 140, (2020), 110212.

[28]. Tomar A., Gupta N., "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures", Science of The Total Environment, 728, (2020), 138762.

[29]. Arora P., Kumar H., Panigrahi B.K., "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India", Chaos, Solitons & Fractals, 139, (2020), 110017.

[30]. Alzahrani S.I., Aljamaan I.A., Al-Fakih E.A., "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions", Journal of Infection and Public Health, 13(7), (2020), 914-919.

[31]. Ribeiro M.H.D.M., et al., "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil", Chaos, Solitons & Fractals, 135, (2020), 109853.

[32]. Sujath R., Chatterjee J.M., Hassanien A.E., "A machine learning forecasting model for COVID-19 pandemic in India", Stochastic Environmental Research and Risk Assessment, (2020).

[33]. Tuli S., et al., "Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing", Internet of Things, 11, (2020), 100222.

[34]. Wang P., et al., "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics", Chaos, Solitons & Fractals, 139, (2020), 110058.

[35]. Liu J., et al., "Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China", Science of The Total Environment, 726, (2020), 138513.

[36]. Goswami K., Bharali S., Hazarika J., "Projections for COVID-19 pandemic in India and effect of temperature and humidity", Diabetes & Metabolic Syndrome: Clinical Research & Reviews, (2020).

[37]. Li H., et al., "Air Pollution and temperature are associated with increased COVID-19 incidence: a time series study", International Journal of Infectious Diseases, (2020).

[38]. Wu Y., et al., "Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries", Science of The Total Environment, 729, (2020), 139051.

[39]. Farzan A.N., U.S. coronavirus cases surpass 1.8 million as concern over potential spread rises with turmoil, in The Washington Post. 2020, The Washington Post.