ICIC Express Letters
Part B: Applications
Volume **12**, Number **2**, February **2021**
 ISSN 2185-2766
pp. **143–151**

FORECASTING THE CUMULATIVE NUMBER OF CONFIRMED CASES OF COVID-19 IN USA, SAUDI ARABIA, AND CHINA USING ENSEMBLE MODEL

Marwan Albahar[1], Abdulaziz Albahr[2,3], Ayman Alharbi[4]
Mohammed Thanoon[1] and Sami Karali[1]

[1]Department of Computer Science
[4]Department of Computer Engineering
Umm Al Qura University
Mecca, PO Box 715, Saudi Arabia
{ mabahar; aarharbi; sfkarali }@uqu.edu.sa

[2]College of Applied Medical Sciences
[3]King Abdullah International Medical Research Center
King Saud Bin Abdulaziz for Health Science
Alahsa Campus, Saudi Arabia
aalbahr400@gmail.com

Received July 2020; accepted September 2020

Abstract. *In December 2019, the novel coronavirus (COVID-19) was spread in Wuhan city, China, which caused an outbreak of respiratory illness. It has been a curiosity for how and how long the number of cases will increase. This study aims to forecast the number of confirmed COVID-19 cases in Saudi Arabia, United States, and China. In this paper, an ensemble Machine Learning (ML) model is proposed for COVID-19 outbreak prediction. A public dataset is used to test the ML model's prediction performance. The prediction performance is measured using the calculation of Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and R squared values. Based on the experiment in this paper, the results show that random forest provides satisfactory prediction performance for these countries with minimum RMSE and MAE values.*
Keywords: COVID-19, Saudi Arabia, USA, China, Forecasting, Ensemble model, Cumulative number

1. **Introduction.** The COVID-19 was spread in China started in Wuhan city in December 2019. COVID-19 made an outbreak of respiratory illness. Nowadays, the study of COVID-19 virus's behavior has devoted considerable attention. Several models have been developed to analyze the outbreak of COVID-19 employing mathematical, dynamic, and statistical techniques [1, 2, 3, 4, 5]. These techniques assisted in estimating the influence of intervention strategies and predicting the dynamics of transmission. Nevertheless, they experienced impediments that lay on the design and the dependency on numerous assumptions. As a consequence, the prediction accuracy of future cases of COVID-19 may not be very accurate. If the spread of COVID-19 is not controlled, and the numbers of infections are still raised, these would overburden the country's healthcare system in the coming days. Also, as an expectation for discovering a vaccination of the COVID-19, it would take a long period because the clinical trials have to be done extensively to avoid higher risks of failure. Therefore, most of the countries took country-wide lockdowns and precautionary measures to control the spread of COVID-19.

An international public health emergency has been reported about 2.1 million confirmed COVID-19 cases and 146 thousand deaths worldwide. China, Middle East, and USA

DOI: 10.24507/icicelb.12.02.143

have been the center of this outbreak. USA is from the most affected countries by this outbreak [6]. Saudi Arabia has the least significant number of cumulative confirmed cases of COVID-19 between Jan 22 and May 31, with 8535 cases. While USA ranks first with the most cumulative confirmed cases in the same period with 109259 cases (See Figure 1) [7]. In the last few months, several studies are done to predict the spread of the disease. Machine Learning (ML) has been used to build spread prediction models because of the current crisis and the global nature of the issue in developing epidemiological models. The approach of ML in developing models gives the ability and utmost prediction accuracy for longer lead-times [8].

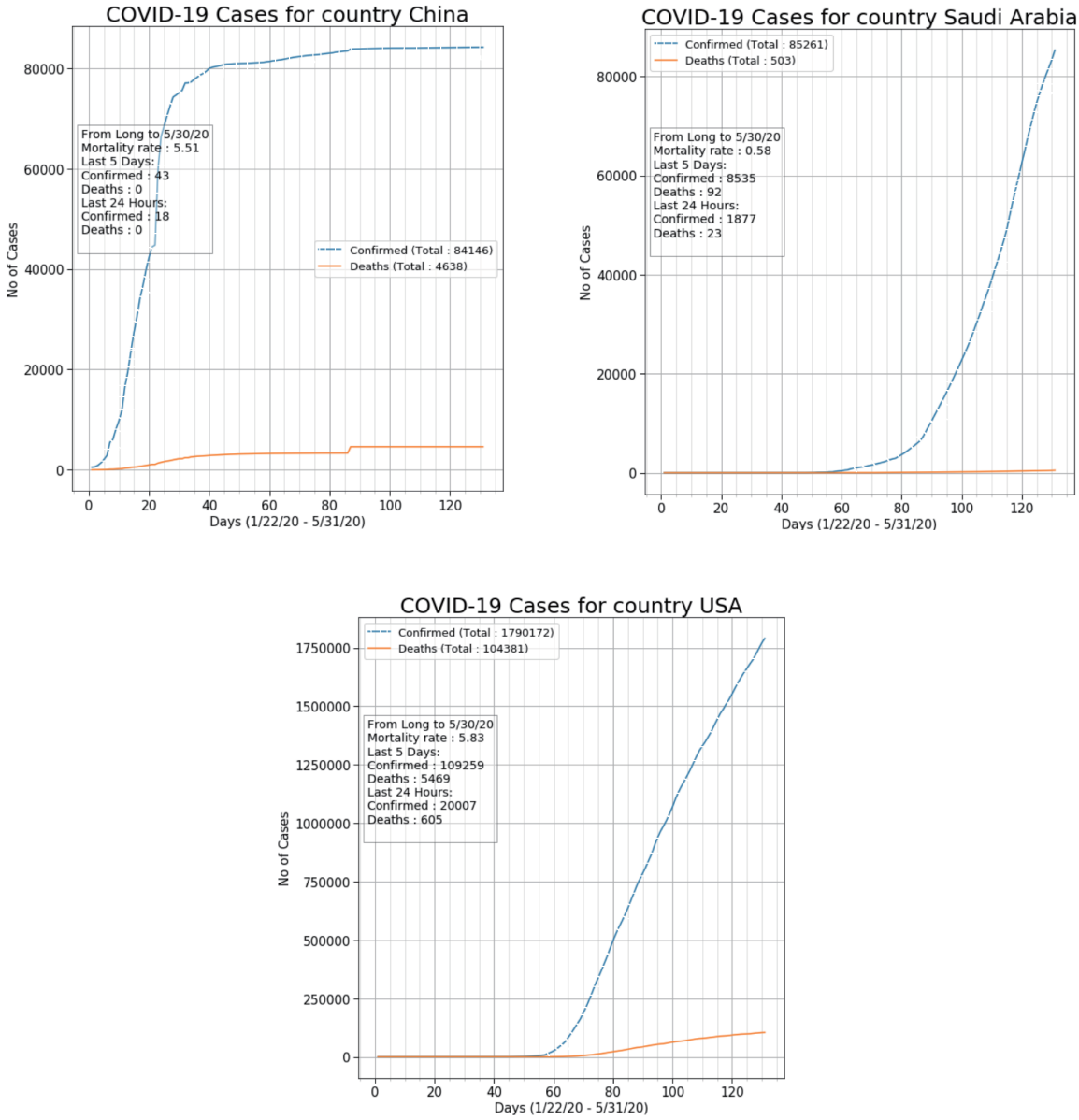

FIGURE 1. Timeline of confirmed/death events between Jan 22 and May 31, 2020

Zhao et al. utilized a mathematical model to estimate the number of COVID-19 cases that had not reported from 1 to 15 January 2020. They concluded that 469 cases were unreported. Also, they estimated the basic reproduction number (R_0) to be 2.56 [3]. In a different work, Nishiura et al. employed a statistical model to predict the ascertainment rate of infected individuals in Wuhan using a dataset collected from Japanese citizens evacuated from Wuhan [9]. In addition, Tang et al. used a mathematical model, called

Susceptible Exposed Infectious Recovered (SEIR), to estimate the transmission rate of COVID-19 and examine the effectiveness of different preventive strategies such as travel restriction on the infections. They found that the basic reproduction number was 6.47 and travel restriction strategy could significantly lower the number of infected people by almost 90% [10]. Thompson estimated the probability of sustained human-to-human transmission in a new location using data of 47 patients. He found that the risk of sustained transmission was 0.41 [11]. Jung et al. proposed a statistical model to predict the risk of death from COVID-19 for two designated scenarios. The estimated risk of death for the two scenarios were 5.1% and 8.4% respectively [12]. Roosa et al. employed three validated phenomenological models during previous outbreaks to create short-time prediction of the total number of confirmed COVID-19 cases [13]. Yang et al. utilized a modified Susceptible Exposed Infectious Recovered (SEIR) model and AI-based model to predict the peak and the size of COVID-19 outbreak. Their estimated results showed that the peak of the COVID-19 epidemic occurred on February 4 for the AI-based model and February 20 for the modified SEIR model [14].

However, such prediction models have encountered many challenges in making predictions in real time such as the daily infection number, basic reproduction number, and virus incubation period. This leads to an inaccurate trend prediction of epidemics in different regions. Therefore, it may be a vital research topic to estimate how far the outbreak would spread and compare the predictive capacity of ensemble learning in the task of forecasting COVID-19 cumulative cases.

In this paper, we mainly focus on three countries that represent a *low-medium-high* daily rate. Time series prediction for some days ahead for COVID-19's cumulative confirmed cases are evaluated using ensemble learning, including the XGBoost algorithm, PLS algorithm, linear SVR algorithm, random forest algorithm. The prime contributions of this paper are as the following.

- We propose an ensemble model used to forecast the number of COVID-19 cases in three countries.
- The outcome of this study is expected to assist governmental initiatives that affect the spread of COVID-19.
- Furthermore, this work is guiding and alternative studies for estimating COVID-19 cases numbers for other countries or provinces.

The paper proceeds as follows. In Section 2, we briefly describe the four algorithms used in this work. In Section 3, we introduce the dataset utilized in this study. In Section 4, we explain our method in detail. Section 5 shows our experimental results, and Section 6 concludes and gives our future works.

2. **Background.**

2.1. **XGBoost algorithm.**

XGBoost is a group tree-boosting process. The Generalized Boosted Model (GBM) algorithm lacks a robust regularization factor because it has been liable to overfitting. Therefore, XGBoost is a new implementation of GBM that provides a robust regularization framework and overcomes overfitting. The new implementation gives XGBoost much popularity these days and has become a state-of-the-art machine learning algorithm. In fact, because of XGBoost's recursive tree-based decision method and the benefit of excellent interpretability potential, XGBoost is classified to be a high-performance machine learning algorithm. Furthermore, XGBoost's accumulated use in every decision stage in trees determines the significance of each individual characteristic.

2.2. **Random forest.**

The Random Forest (RF) algorithm is a group learning process paired with various decision tree predictors. Each decision tree predictor is trained by RFs independently using feature subsets and random data samples. The randomness of

data supports RF to not be a single decision tree but to be more robust and less potential to overfit on the training data.

2.3. **Partial Least Squares (PLS) regression algorithm.** Partial Least Squares (PLS) regression algorithm is a statistical method that bears several relevances to the main components regression. The algorithm discovers a linear regression model through projecting the observable variables and predicted variables to a new space instead of discovering maximum variance's hyperplanes between independent variables and the response. Between two matrices (X and Y), the linear covariance structure is modeled by PLS using the projection to a potential space approach. A PLS attempts to discover the multidimensional direction in the X space that clarifies the maximum multidimensional alteration direction in the Y space.

2.4. **Linear support vector regression algorithm.** The Support Vector Regression (SVR) and Support Vector Machine (SVM) utilize the same essentials for classification with a slight minor difference. SVR has a problem which is to discover a function based on a training sample that approximates an input domain to real numbers. Therefore, a decision boundary at a distance from the premier hyperplane is the prime aim that needs to be decided to be within the boundary line.

3. **Dataset.** We utilize public data, which is served by the Johns Hopkins University Centre for Systems Science and Engineering (JHU CSSE), to create a dataset that includes a combination of several cases [15]. The repository consists of provinces or states of each country where COVID-19 cases are confirmed. Firstly, We observed the five records of the confirmed cases. Secondly, we looked for missing values. Lastly, for this research, we chose USA, China, and Saudi Arabia to examine COVID-19 outbreak.

4. **Proposed Model.**

4.1. **Pre processing and assignment.** Here, we started by changing the name of the features. Thus, the Province/State features became a state. The country is for the Country Region feature. Finally, the name of the countries was changed by using pycountry_convert, an extension of the python package called pycountry that provides conversion functions.

4.1.1. *Training and testing.* Studying the new confirmed COVID-19 cases data for each day helps us find the most accurate fitting distribution models. Therefore, five sets from the daily new confirmed cases of COVID-19 data were chosen to fit the different types of distributions parameters. Following that, the best five performing distributions were recognized. In this model, we trained and combined 200 trees predictions for each date of confirmed cases. After getting 200 predictions of each period, we added all the predictions and got the average forecast of all 200 predicted values. In this way, we got the robust case predictions after getting 200 predictions of each day prediction. In the training process, we used all the dates from 01/22/2020 to 5/31/2020 (**discarded 22nd Jan 2020 because, on this day, all the counts of confirmed cases are zero**). We tuned the parameters of each algorithm as well for better results. In the testing process, we tested all the data on the trained model, including zero's cases (**like in the USA and Saudi Arabia, first 50 days the cases were zero**). Still, we considered these days to check the performance of the model with predictions. We calculated the actual cases and predictions using statistical functions to monitor the performance. Figure 2 presents the architecture of our proposed model.

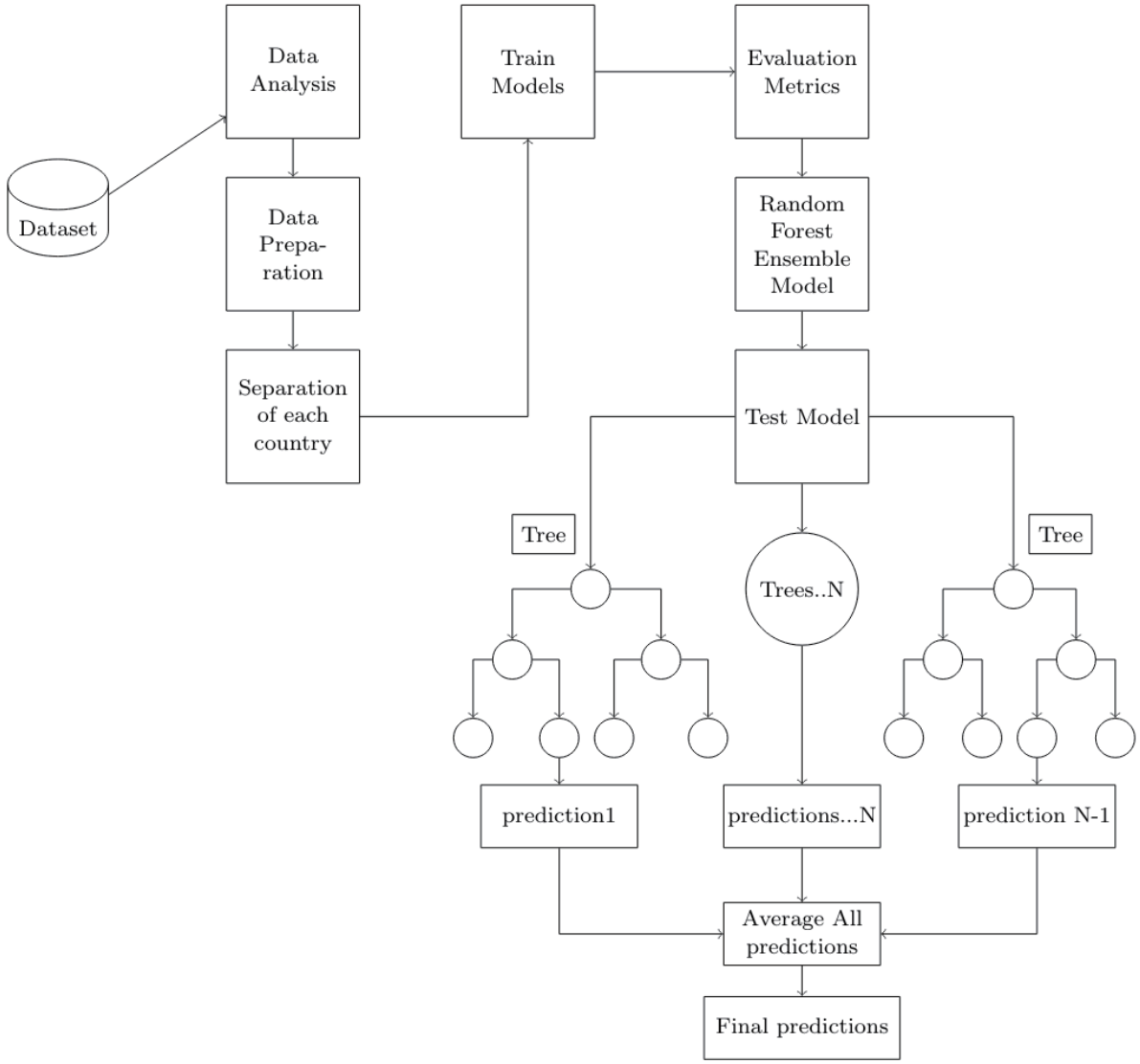

FIGURE 2. Our proposed model architecture

4.1.2. *Hyperparameters setting.* We tune the hyperparameters of each algorithm to get the optimal solution/predictions. Hyperparameter tuning is essential because these control the behavior of the machine learning model. We tune hyperparameter of all algorithms, we yield the best results on RF regression algorithm with the following settings of parameter: n_estimators = 200, min_impurity_decrease = 0.8, min_samples_split = 2, criterion = 'mae'.

5. **Result.** The preliminary outcomes of the cumulative confirmed cases of COVID-19 were compared for China, Saudi Arabia, and USA. The World Health Organization's database was used to take the case numbers. The evaluation of the comparison was done by using R squared, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The calculations of each machine learning algorithm for USA, China, and Saudi Arabia are shown in Table 1, Table 2, and Table 3, respectively.

Investigating the above tables confirms that XGB, PLS, and linear SVR algorithms are poorly performing and misleading information about the actual and predicted cases. On the other hand, RF regression performed well, with the minimum root mean squared

error values and mean absolute error values. That is because RF regression algorithm is an ensemble and works well in the bagging techniques. The RF trees are run in a parallel fashion. The ensemble modeling algorithm selects the random predictions by averaging each random tree prediction. Consequently, we got as close predictions as the true cases in USA, China, and Saudi Arabia, as shown in Figures 3, 4, and 5, respectively.

TABLE 1. USA confirmed cases

Algorithms	**Evaluation metrics**			
	R Squared	MAE	RMSE	MAPE
XGBOOST	0.9935	359.0164	1040.1940	473.0454
PLS	0.668	5892.3495	7449.651	4855.751
Linear SVR	−0.310	11901.678	14807.966	41661.4747
Random forest regression	0.994	214.26	974.44	271.51

TABLE 2. China confirmed cases

Algorithms	**Evaluation metrics**			
	R Squared	MAE	RMSE	MAPE
XGBOOST	0.839	87.57	674.56	9.179
PLS	0.224	587.667	1482.309	28.246
Linear SVR	−0.019	290.942	1699.047	21.492
Random forest regression	0.884	15.730	573.187	0.81

TABLE 3. Saudi Arabia confirmed cases

Algorithms	**Evaluation metrics**			
	R Squared	MAE	RMSE	MAPE
XGBOOST	0.99	15.105	36.78	0.32
PLS	0.74	371.38	434.359	9.87
Linear SVR	−0.78	1311.56	1155.99	52.58
Random forest regression	0.99	11.84	39.17	0.208

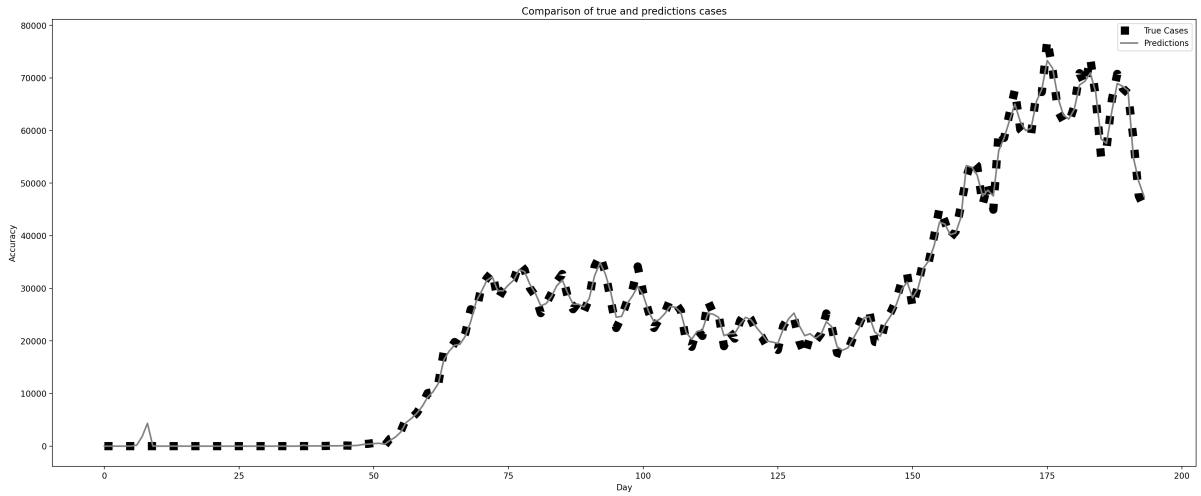

FIGURE 3. Prediction and forecasting results of the cumulative cases of COVID-19 for the USA using RF

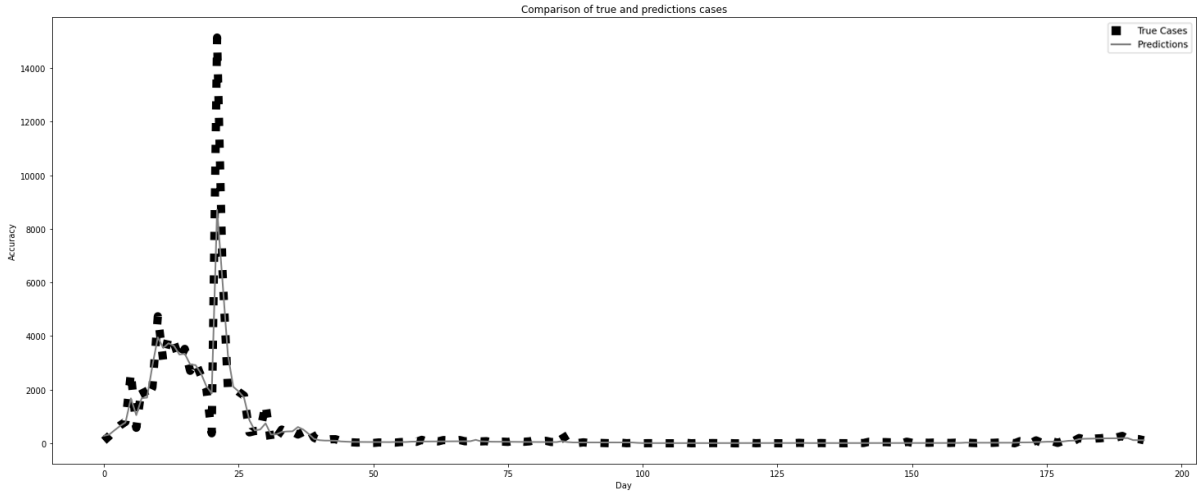

FIGURE 4. Prediction and forecasting results of the cumulative cases of COVID-19 for the China using RF

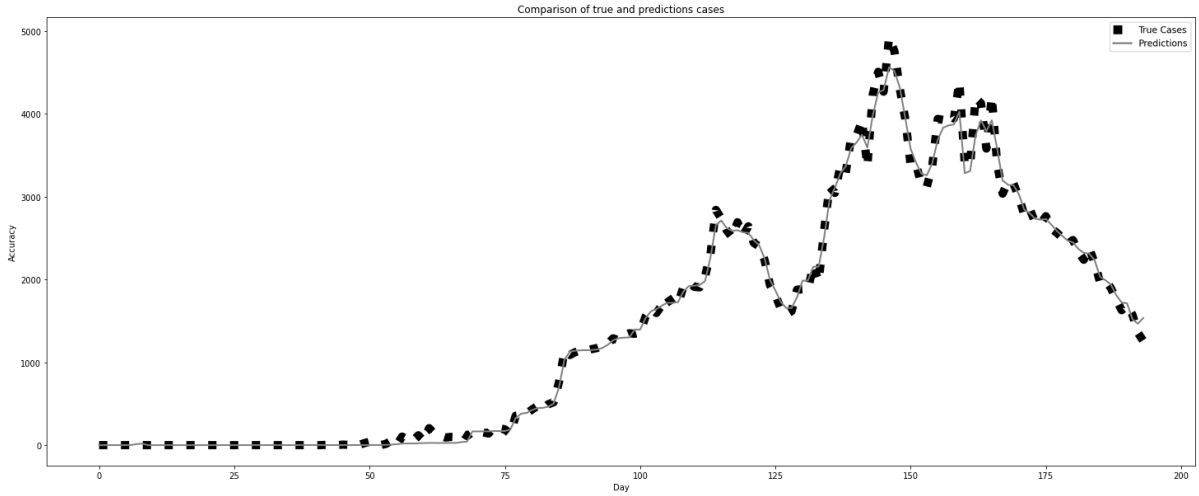

FIGURE 5. Prediction and forecasting results of the cumulative cases of COVID-19 for the Saudi Arabia using RF

6. **Conclusion.** This research study proposes predicted results of the COVID-19's cumulative case numbers in China, Saudi Arabia, USA. Four algorithms were utilized in the proposed research that includes: XGBoost, PLS, linear SVR, and random forest algorithms. These algorithms were used to forecast the case numbers from 01/22/2020 to 5/31/2020. The evaluation of these algorithms was tested by using R squared, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) techniques. The proposed research concludes the following:

- RF algorithm produces the most accurate prediction results with having the smallest root mean squared error and mean absolute error.
- The COVID-19's cumulative case numbers in China, Saudi Arabia, and USA are forecasted as approximately 12372, 11299, and 17, respectively, from 06/02/2020 to 06/08/2020 by using RF (See Figures 6-8).
- It is estimated that the cumulative cases number of COVID-19 will grow at a diminishing rate until June 08, 2020, for USA, Saudi Arabia.

For future works, we intend to empower the prediction process by taking account of government-initiated precautionary measures to produce meaningful analysis and possibly precise predictions. So, we will investigate the capacity to incorporate the precautionary

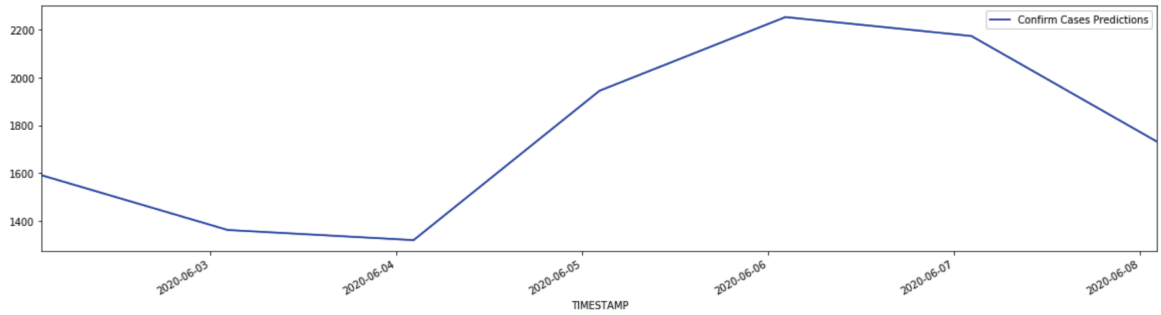

FIGURE 6. Prediction results of the cumulative cases of COVID-19 in USA for the period from 06/02/2020 to 06/08/2020 using RF

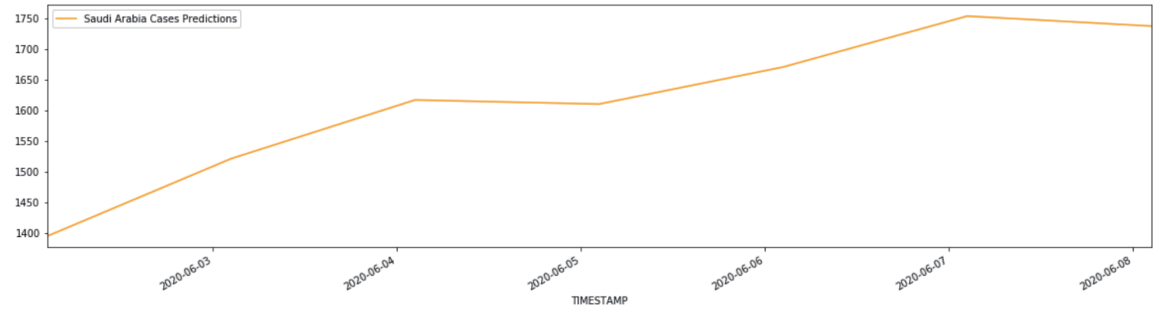

FIGURE 7. Prediction results of the cumulative cases of COVID-19 in Saudi Arabia for the period from 06/02/2020 to 06/08/2020 using RF

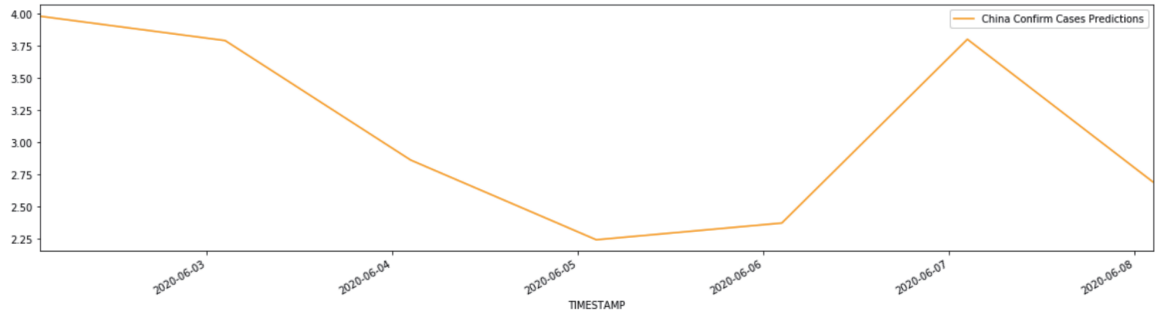

FIGURE 8. Prediction results of the cumulative cases of COVID-19 in China for the period from 06/02/2020 to 06/08/2020 using RF

action taken by governments **(as changing interventions in reality)** aiming to forecast the future trend of COVID-19 spread.

REFERENCES

[1] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung and Z. Feng, Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia, *New England Journal of Medicine*, 2020.

[2] J. T. Wu, K. Leung and G. M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study, *The Lancet*, vol.395, pp.689-697, 2020.

[3] S. Zhao, S. S. Musa, Q. Lin, J. Ran, G. Yang, W. Wang, Y. Lou, L. Yang, D. Gao, D. He and M. H. Wang, Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China

in the first half of January 2020: A data-driven modelling analysis of the early outbreak, *Journal of Clinical Medicine*, vol.9, p.388, 2020.

[4] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, N. Davies, A. Gimma, K. van Zandvoort, H. Gibbs, J. Hellewell, C. I. Jarvis, S. Clifford, B. J. Quilty, N. I. Bosse, S. Abbott, P. Klepac and S. Flasche, Early dynamics of transmission and control of COVID-19: A mathematical modelling study, *The Lancet Infectious Diseases*, vol.20, pp.553-558, 2020.

[5] A. R. Tuite and D. N. Fisman, Reporting, epidemic growth, and reproduction numbers for the 2019 novel coronavirus (2019-nCoV) epidemic, *Annals of Internal Medicine*, 2020.

[6] W. WHO, 2020a, *Coronavirus Disease (COVID-19) – Events as They Happen*, 2020.

[7] W. WHO, 2020a, *Who Coronavirus Disease (COVID-19) Dashboard*, 2020.

[8] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. Varkonyi-Koczy, U. Reuter, T. Rabczuk and P. Atkinson, COVID-19 outbreak prediction with machine learning, *Preprints*, DOI: 10.20944/preprints202004.0311.v1, 2020.

[9] H. Nishiura, T. Kobayashi, Y. Yang, K. Hayashi, T. Miyama, R. Kinoshita, N. Linton, S.-M. Jung, B. Yuan, A. Suzuki and A. Akhmetzhanov, The rate of underascertainment of novel coronavirus (2019-nCoV) infection: Estimation using Japanese passengers data on evacuation flights, *Journal of Clinical Medicine*, vol.9, p.419, 2020.

[10] B. Tang, X. Wang, Q. Li, N. Bragazzi, S. Tang, Y. Xiao and J. Wu, Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions, *Journal of Clinical Medicine*, 2020.

[11] R. Thompson, Novel coronavirus outbreak in Wuhan, China, 2020: Intense surveillance is vital for preventing sustained transmission in new locations, *Journal of Clinical Medicine*, vol.9, p.498, 2020.

[12] S.-M. Jung, A. R. Akhmetzhanov, K. Hayashi, N. M. Linton, Y. Yang, B. Yuan, T. Kobayashi, R. Kinoshita and H. Nishiura, Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases, *Journal of Clinical Medicine*, vol.9, p.523, 2020.

[13] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. Hyman, P. Yan and G. Chowell, Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020, *Infectious Disease Modelling*, vol.5, pp.256-263, 2020.

[14] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong and J. He, Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, *Journal of Thoracic Disease*, vol.3, pp.165-174, 2020.

[15] E. Dong, H. Du and L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, *The Lancet Infectious Diseases*, vol.20, pp.533-534, 2020.