A Novel Machine Learning based Model for COVID-19 Prediction

Tamer Sh. Mazen
Lecturer Dept. Management Information System
Modern Academy of Computer Science and Management, Cairo, Egypt

***Abstract*—Since end of 2019, the World Health Organization (WHO) provided the name COVID-19 for the disease caused by the novel coronavirus. Coronavirus is a family of viruses that are named according to the spiky crown existed on the outer surface of the virus. The novel coronavirus, also known as SARS-CoV-2, which is a contagious respiratory virus that first reported in Wuhan, China. According to the rapid and sudden spread for COVID-19, it attracts most of the scientists and researchers all over the world. Researchers in the data science field are trying to analyze the worldwide infection cases day-by-day to gain a complete statistical view of the current situation. In this paper, a novel approach to predict the daily infection records for COVID-19 is presented. The model is applied for Egypt as well as the highest 10 ranked countries based on the number of cases and rate of change. The proposed model is implemented based on supervised Machine-Learning Regression algorithms. The dataset used for prediction was issued by WHO starting from 22 Jan 2020.**

***Keywords*—Coronavirus; COVID-19; coronavirus in Egypt; supervised machine learning; regression models**

I. Introduction

Since the end of 2019, the outbreak of COVID-19 began in Wuhan, China. The new virus is a form of Coronaviruses, that affects the respiratory system such as the SARS virus. COVID-19 consists of a protein membrane with a diameter of 50-200-200 nm, inside which the DNA of the RNA virus is enveloped, which forms the spinal bumps on the surface of the virus and gives it a distinctive coronary shape [1], [2]. “Fig. 1”, shows the internal structure of SARS-COVID virus [3].

Rational decisions are the goal that governments seek so as to address the COVID-19 epidemic. The prediction process is one of the most important tools needed to face that problem. The prediction models are used to predict the number of new daily confirmed cases, recoveries, and deaths. The prediction of newly confirmed cases helps the governments to update their precautionary procedures as well as getting ready by the needed hospitals equipment and the human preparation.

Nowadays, the main target is to find a cure for the killing virus as well as to predict its spread rate. Many researches in the data science field were found to study the statistical situation of the virus.

Furqan Rustam et al. [4] tested a set of different Machine Learning based models in order to predict the number of future COVID-19 patients. The used models are linear regression, least-absolute shrinkage and selection operator, support vector machines, and exponential smoothing. Results showed that the exponential smoothing based model provided the most accurate prediction results, while the support vector machines provided the worst results as compared to the four selected models.

Nanning Zheng et al. [5] proposed a new hybrid Artificial Intelligence (AI) based model using Natural Language Processing (NLP), and the Long-Short Term Memory (LSTM) network, in addition to the Improved-Susceptible Infected (ISI) model, presented so as to predict the future cases in China. The proposed model could predict to a high degree the actual epidemic-cases.

In [6] Li Yan et al. proposed a machine learning based model using 3 clinical parameters to detect the new death rate of current patients. The accuracy of the proposed model is more than 90%.

Renato R. Silva et al. [7] used a Bayesian based methodology in order to detect the peak of the outbreak in one of the Brazilian countries (Goias) based on the number of confirmed cases. They found that, the peak will be reached between 7 to 10 weeks from the beginning of the crisis supposing that, there will be no change in the governmental control during the upcoming period.

In this paper, a novel prediction model that predicts the number of new confirmed cases is presented. The proposed model uses a set of statistical based techniques in a supervised machine learning process. The model is tested on Egypt as well as the top 10 ranked countries for COVID-19 till end of September 2020. The results of the proposed model are compared against the Bayesian Ridge regression model.

The next sections of the paper will be as follows. In Section 2, the distribution of COVID-19 all over the world is presented. Section 3, shows COVID-19 in Egypt. The proposed model is explained in Section 4 followed by the experimental results in Section 5. Finally, the conclusion is presented in Section 6.

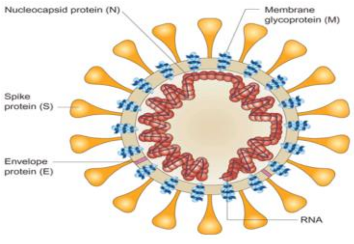


Fig. 1. SARS-COVID Structure.

II. State of the Art

As a subset of Data Science, Artificial Intelligence (AI) and Machine Learning (ML) are playing a major role in the analysis and visualization of the COVID-19 crisis. These predictions will provide a help to the healthcare systems and government institutions to speed up investigations about the virus rapid and terrible spread.

"Fig. 2" illustrates the distribution for COVID -19 case all over the world. During the interval starting from 22 January till end of September 2020, the number of confirmed cases, deaths, and recovery cases are (38,917,803), (1,098,254), and (26,885,286) consequently.

"Table I" and "Fig. 3", show the number of confirmed cases, deaths and recovery cases over the world grouped by continents. As seen, Europe is the most affected continent followed by Asia, America, Africa and finally Australia.

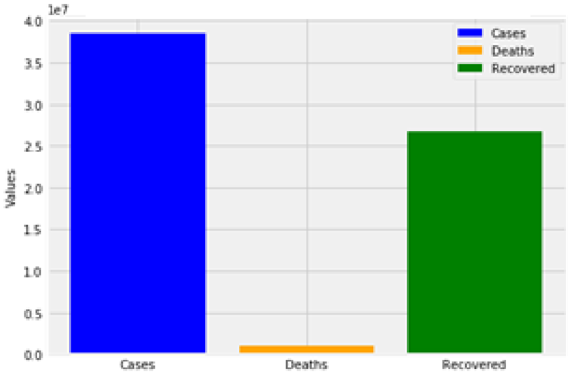


Fig. 2. COVID-19 Distribution All Over the World.

TABLE I. COVID-19 Confirmed, Death and Recovered Case Over the World Till 30 September 2020

Continent	Confirmed	Deaths	Recovered
Africa	10450293	38407	1285804
America	4672236	384875	8993415
Asia	13986946	214996	10341346
Europe	8584922	238143	2867398
Australia	748801	938	27438

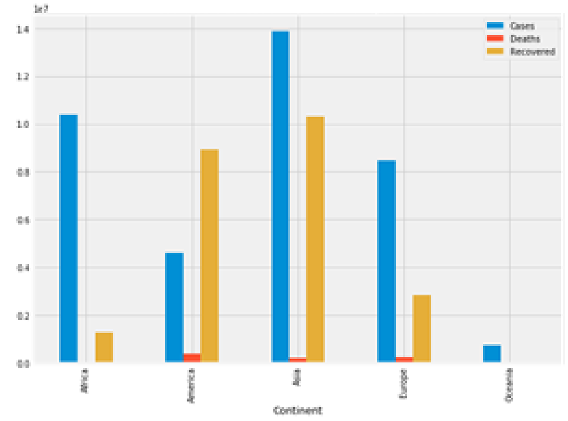


Fig. 3. COVID-19 Distribution over the World/Continent.

"Table II" and "Fig. 4", show the number of confirmed cases, deaths, and recovery cases for the mostly infected 10 countries classified by WHO on 30 September. "Table III" and "Fig. 5", show the rate of change percentage of the top 10 countries.

TABLE II. Top 10 Countries based on COVID-19 Confirmed Cases

Country	Confirmed	Deaths	Recovered
USA	7916099	216872	3155794
India	7307097	111266	6383441
Brazil	5140863	151747	4526393
Russia	1332824	23069	1035141
Argentina	931967	24921	751146
Colombia	930159	28306	816667
Spain	908056	33413	150376
Peru	853974	33419	753959
Mexico	829396	84898	703489
France	820376	33058	106374

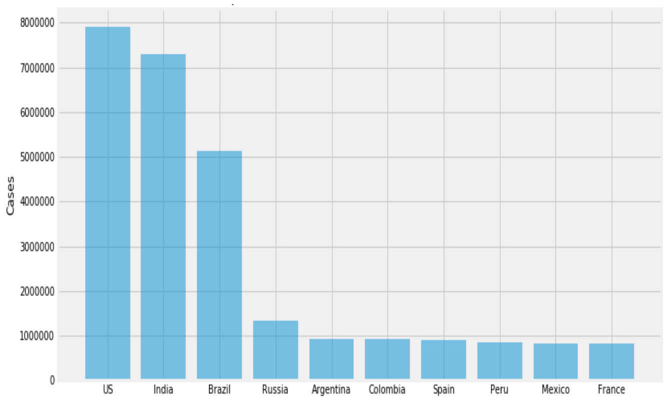


Fig. 4. Top 10 Countries on Deaths Cases in COVID-19.

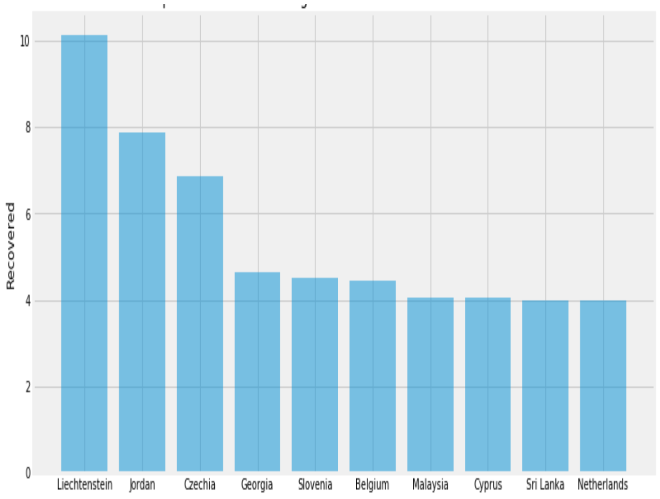


Fig. 5. Top 10 COVID-19 Countries based on Rate of Change.

TABLE III. TOP 10 COVID-19 COUNTRIES BASED ON RATE OF CHANGE

Country/Region	**Rate of change %**
Liechtenstein	10.14
Jordan	7.88
Czechia	6.86
Georgia	4.64
Slovenia	4.52
Belgium	4.44
Malaysia	4.07
Cyprus	4.05
Sri Lanka	4.00
Netherlands	3.98

III. COVID-19 IN EGYPT

In this paper, the rank of Egypt based on the number of confirmed cases and the rate of change is calculated. It is found that, Egypt's rank is 43 around the world based on the number of infections. while, its rank based on the change in rate is 143. These calculations are performed using the WHO dataset till end of September 2020.

"Table IV" and "Fig. 6" show COVID-19 cases in Egypt till 30 September 2020.

There are different regression models found in the literature to predict the number of new confirmed COVID-19 cases such as Support Vector Machines (SVM) [8], Linear regression [9], binomial regression [10], and Bayesian Ridge regression [11].

During our experiments, it was found that the Bayesian Ridge regression model could provide the most accurate prediction results as compared to the other techniques.

TABLE IV. COVID-19 CONFIRMED, DEATH AND RECOVERY CASES IN EGYPT

Country	***Confirmed***	***Deaths***	***Recovered***
Egypt	*104915*	*6077*	*97920*

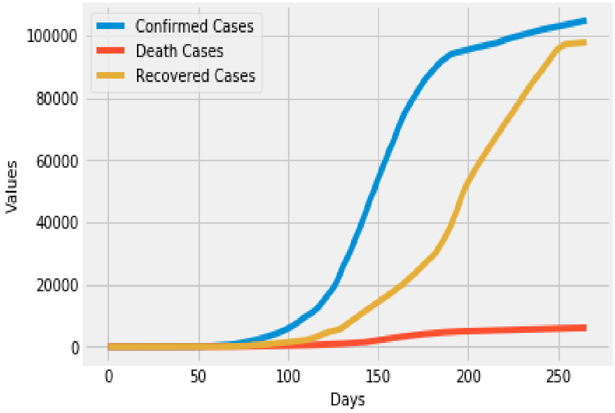


Fig. 6. COVID-19 Cases in Egypt.

IV. PROPOSED PREDICTION MODEL

The main idea of this paper is to build a hybrid model based on mathematical and statistical approaches in a machine learning based environment. The model is performed using rate of change, geometric mean and standard deviation [12] [13] [14] [15].

As seen in "Fig. 7", the data set consisting of the number of confirmed COVID-19 cases [16] , number of deaths [17], and the number of recovery cases [18] is collected. Then, the data preprocessing is performed and the data is split into 80% training and 20% testing. The steps of the proposed model are explained in details with an example on Egypt.

"Table V" shows the sample of data used as an example where, X represents the days starting from the 110th to the 119th day from the start of the epidemic while Y represents the number of confirmed cases on each day (each of the numbers below is multiplied by 10^3).

A. *Steps of the Proposed Model*

Step 1: Splitting the data set

Dataset was split into a training set and testing set, 80%, and 20%, respectively.

Step 2: Calculating the number of newly confirmed

Where new cases are calculated as,

$$New_{Case[i]} = y[i] - y[i-1] \quad (1)$$

Where, $y[i]$ is the number of confirmed cases at day i.

Step 3: Calculating the Rate of Change (RoC):

The Rate of Change (RoC) for the newly confirmed cases is calculated by the next formula [19].

"Table VI" shows the RoC calculation for the sample example.

$$RoC = \frac{\boldsymbol{New_{Case}[i]}}{\boldsymbol{y[i-1]}} \quad (2)$$

Step 4: Calculating the Geometric mean (GM)

It is a n^{th} root for the RoC for n days [20] [21].

$$GM = \left(\prod_{i=1}^{n}(RoC^2)\right)^{1/n} \quad (3)$$

Here, in this example the GM = 0.149956677.

TABLE V. SAMPLE OF COLLECTED DATA ABOUT EGYPT

X	***110***	***111***	***112***	***113***	***114***	***115***	***116***	***117***	***118***	***119***
Y	*9.4*	*9.7*	*10*	*10.4*	*10.8*	*11.2*	*11.7*	*12.2*	*12.7*	*13.4*

Step 5: Calculating the standard deviation,

Standard Deviation "STD" is calculating the extent of deviation of values from the average value:

$$STD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2} \qquad (4)$$

Where, n is the number of days.

Here, in this example the STD = 0.1334

Step 5: Calculating the newly expected cases and boundaries

The new expected cases based on the proposed model are calculated using the following formula:

$$Exp_{Case} = present_{newcase} * 1 + GM \qquad (5)$$

$$Exp_{Lower} = Exp_{Case} - (Exp_{Case} - STD) \qquad (6)$$

$$Exp_{Upper} = Exp_{Case} + (Exp_{Case} - STD) \qquad (7)$$

"Table VII" represent results of steps 4, 5 and 6 calculating GM, SD and boundaries.

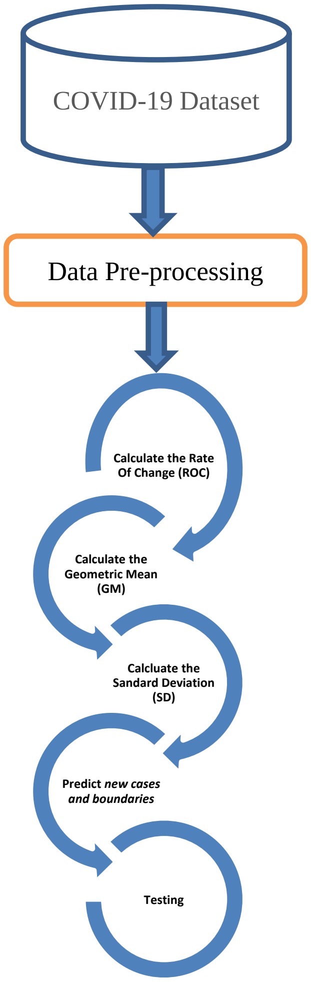

COVID-19 Dataset
Data Pre-processing
Calculate the Rate Of Change (ROC)
Calculate the Geometric Mean (GM)
Calcluate the Sandard Deviation (SD)
Predict new cases and boundaries
Testing

Fig. 7. The Proposed Model Diagram.

TABLE VI. THE ROC CALCULATION

X	110	111	112	113	114	115	116	117	118	119
Y	9.4	9.7	10	10.4	10.8	11.2	11.7	12.2	12.7	13.4
New Case	-	346	347	338	398	399	491	510	535	720
RoC			0.03	0.02	0.17	0.03	0.23	0.08	0.04	0.34

TABLE VII. THE EXPECTED NEW CASES AND BOUNDARIES

X	110	111	112	113	114	115	116	117	118	119
Y	9.4	9.7	10	10.4	10.8	11.2	11.7	12.2	12.7	13.4
New Case	-	346	347	338	398	399	491	510	535	720
RoC			0.03	0.02	0.17	0.03	0.23	0.08	0.04	0.34
Case			348	334	418	415	529	546	571	588
Lower					327	372	376	467	488	516
Upper					341	465	543	591	604	627

B. Testing the Model Accuracy

The proposed model accuracy is tested using both the Mean Square Error (MSE) [22] [23] and the correlation (R) between the expected values and the real values [24] [25].

$$MSE = \frac{\sum_{i=1}^{n}(E^2)}{n} \qquad (8)$$

Where E is the difference between expected and real data.

Here, in the example the $MSE = \frac{33016}{7} = 4716.571429$.

$$r = corr(Exp_{Case} , New_{Case}) \qquad (9)$$

Here, in the example $r = 0.880848$

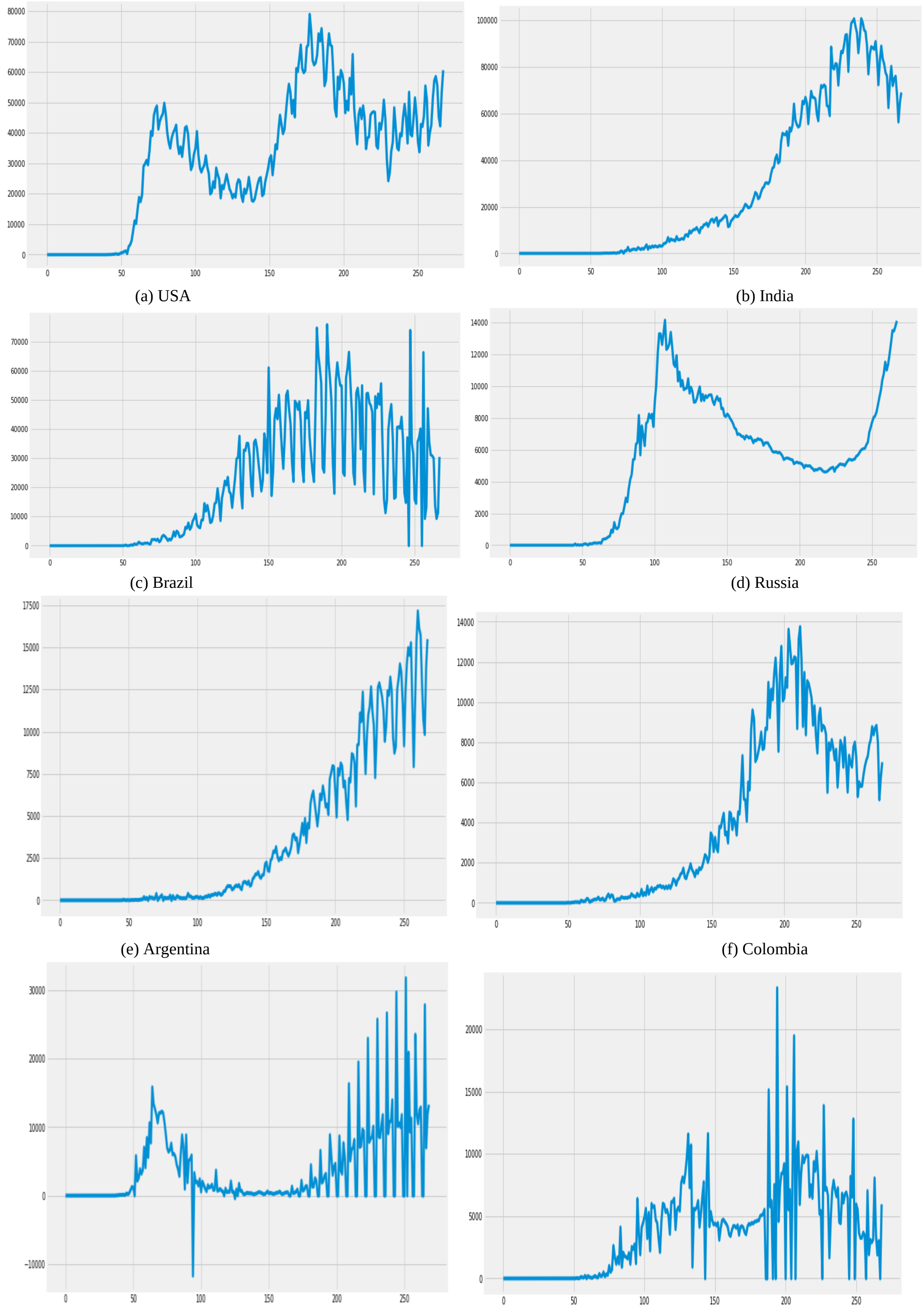
There is a strong relation between results that indicate to model has highest accuracy.

V. EXPERIMENTAL RESULTS

The proposed model is compared against the Bayesian Ridge regression model, as it was most accurate model for COVID-19 predictions amongst the other state of the art techniques.

"Fig. 8" illustrates the daily cases that predicted by proposed model for the highest rated for COVID-19 till end of September 2020. These 10 highest countries are USA, India, Brazil, Russia, Argentina, Colombia, Spain, Peru, Mexico, and France.

"Fig. 9" illustrates the comparison between the proposed model versus the Bayesian Ridge model applied for Egypt as well as the 10 highest rated countries for COVID-19 till end of September 2020. The red lines in the figure represent the daily prediction results while the blue lines represent the real values. As seen, the proposed model is more accurate than its counterpart over all the countries.



(a) USA

(b) India

(c) Brazil

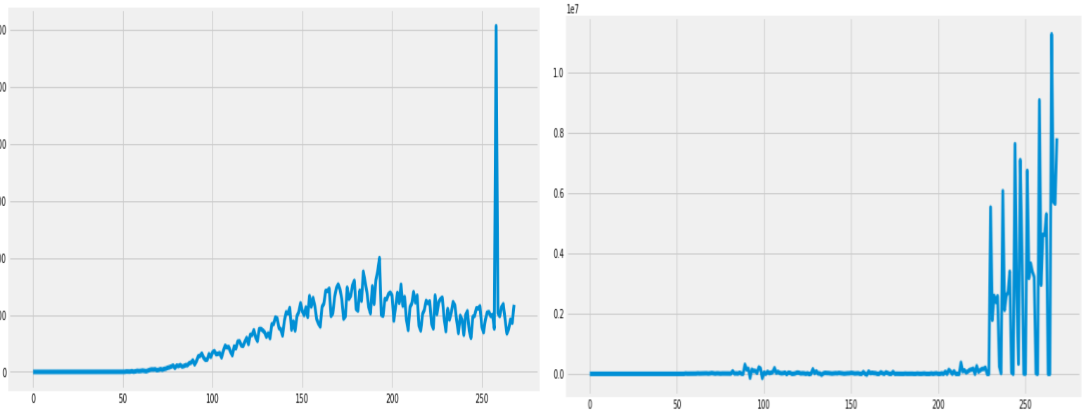
(d) Russia

(e) Argentina

(f) Colombia

(g) Spain

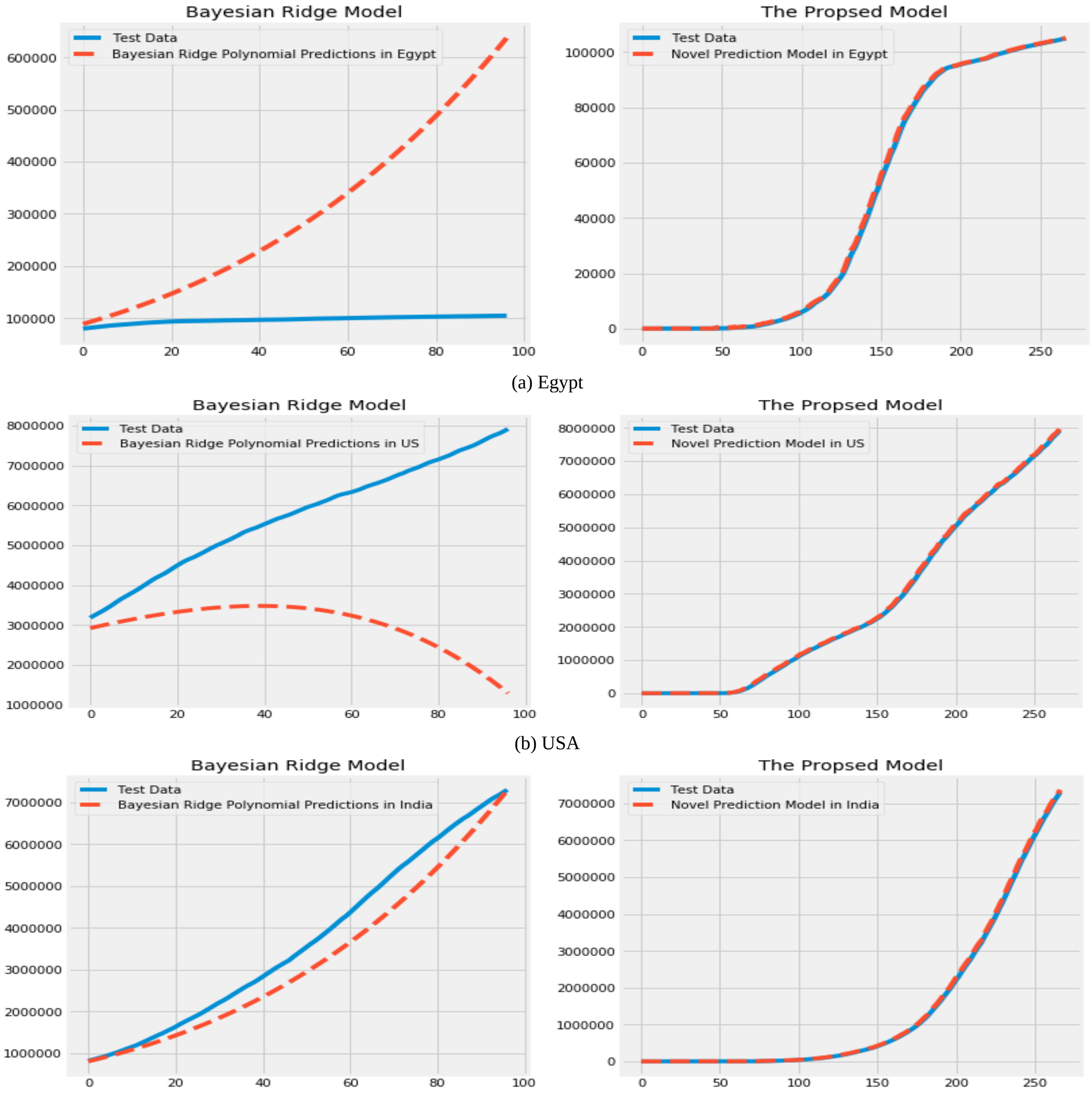
(H) Peru



(i) Mexico

(j) France

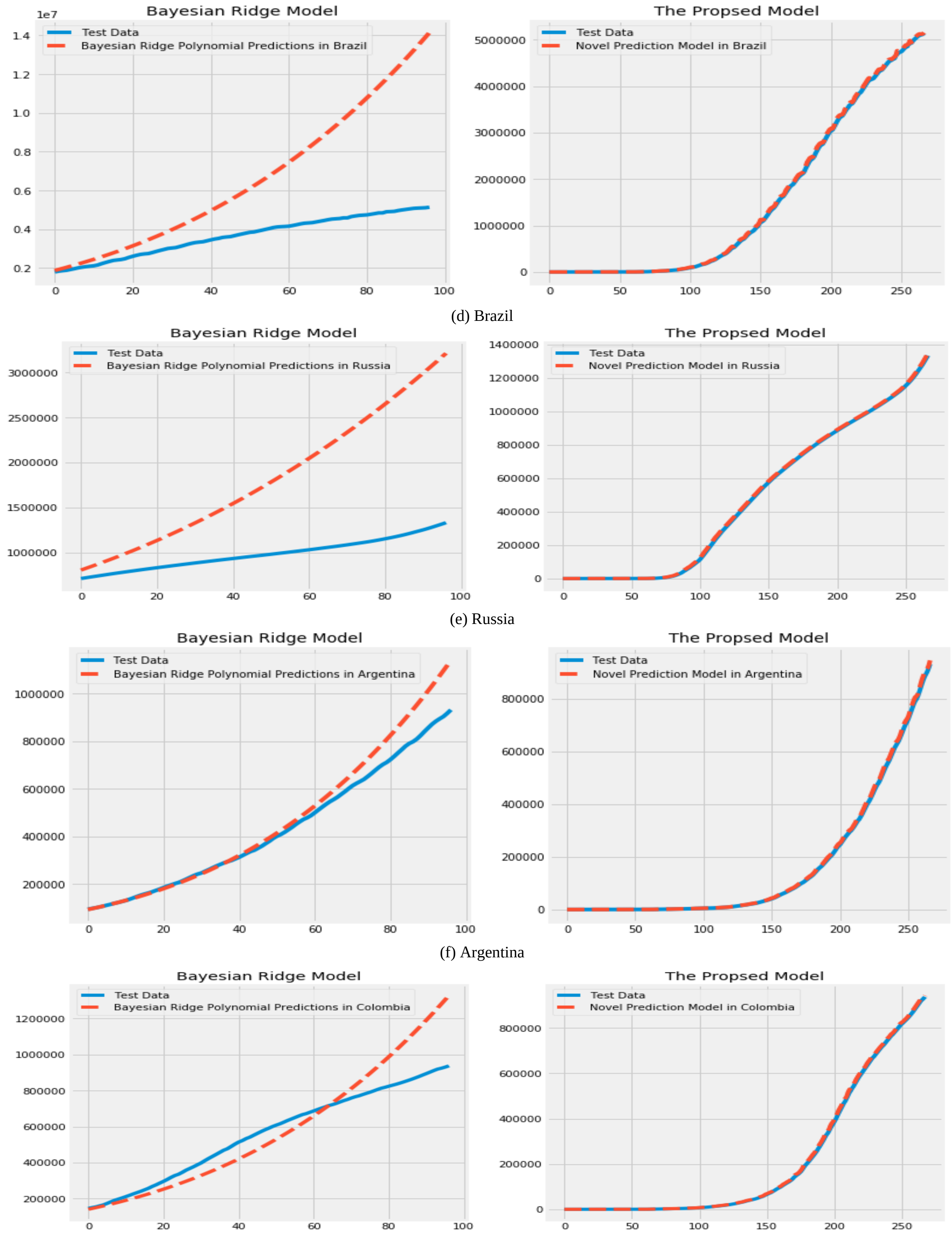
Fig. 8. Daily Prediction Cases used Model.



(a) Egypt

(b) USA

(c) India

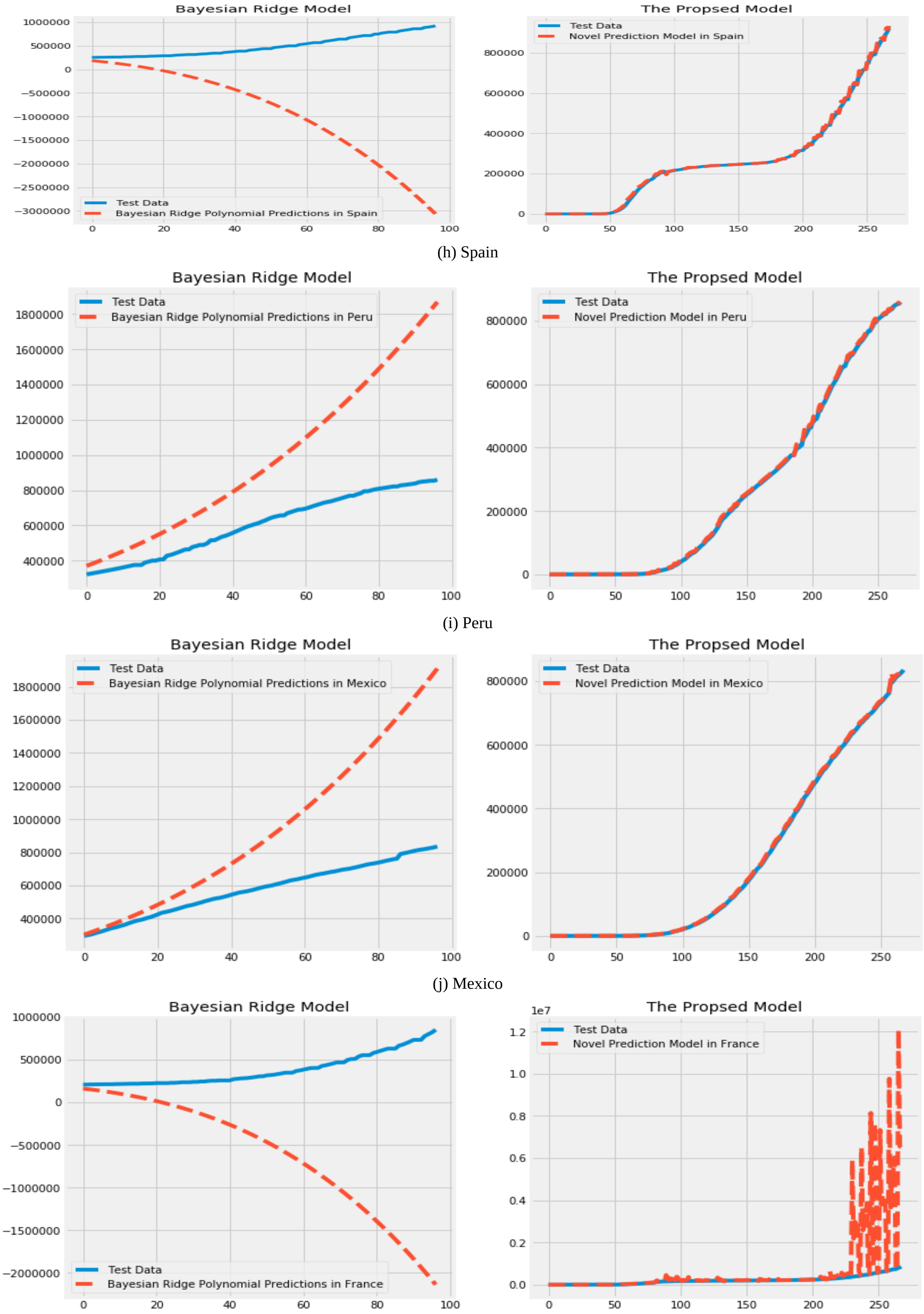


(d) Brazil

(e) Russia

(f) Argentina

(g) Colombia



(h) Spain

(i) Peru

(j) Mexico

(k) France

Fig. 9. Bayesian Ridge Model vs the Proposed Model Daily Predictions for Egypt as well as the Highly Rated 10 Countries.

The sharp peaks found for the prediction results corresponding to France were according to the sudden jump in the number of cases during that period.

The Mean Square Error (MSE) for the two models are presented in “Table VIII” as seen, the MSE for the proposed model over all the testing countries is less than that of Bayesian Ridge model.

TABLE VIII. MODELS EVALUATION COMPARISON MSE

Country	Bayesian Ridge	Proposed Model
	MSE	MSE
Egypt	13.57	11.44
US	29.19	17.03
India	22.74	16.21
Brazil	26.16	20.23
Russia	26.29	15.55
Argentina	19.19	14.73
Colombia	26.18	13.77
Spain	18.52	14.89
Peru	22.76	14.18
Mexico	26.86	15.64
France	26.00	19.29

VI. CONCLUSION

COVID-19 or corona virus pandemic is the danger that threaten both peoples and governments all over the world. Many researches tried to predict the number of newly infected cases, deaths, and recoveries. In this paper, a new hybrid-machine learning based model is proposed so as to predict the newly expected infections. The model is tested on Egypt as well as the 10 highly rated COVID-19 countries till end of September 2020. The proposed model is compared against one of the most accurate prediction models found in the literature i.e. Bayesian Ridge model. Results showed the powerful of the proposed model as compared to its counterpart all over the countries under study.

REFERENCES

[1] Yan Gao et al., "Structure of the RNA-dependent RNA polymerase from COVID-19 virus," Science, vol. 368, no. 6492, pp. 779-782, 15 May 2020.

[2] S. Anastasopoulou and M. Athanasia, "The biology of SARS-CoV-2 and the ensuing COVID-19," ACHAIKI IATRIKI, vol. 39, no. 1, p. :29–35, 2020.

[3] Stadler K, Masignani V, Eickmann M, et al., "SARS--beginning to understand a new virus," Nat. Rev. Microbiol., vol. 1, no. 3, p. 209–218, 2003.

[4] F. Rustam, A. MEHMOOD, A. RESHI , S. ULLAH, B.-W. ON4, W. ASLAM and G. S. CHOI , "COVID-19 Future Forecasting Using Supervised Machine Learning Models," IEEE Access , vol. 8, pp. 101489 - 101499, 25 May 2020.

[5] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, T. Yang, B. Lou, Y. Chi, H. Long, M. Ma, Q. Yuan and S. Zhang, "Predicting COVID-19 in China Using Hybrid AI Model," IEEE Trans Cybern., 8 May 2020.

[6] Yan, L., Zhang, H., Goncalves, J. et al., "An interpretable mortality prediction model for COVID-19 patients," Nature Machine Intelligence, vol. 2, pp. 283-288, 2020.

[7] R. Silva, . W. D. Velasco, d.-S. W. Marques and C. A. G. Tibiica, "A Bayesian analysis of the total number of casesof the COVID 19 when only a few data isavailable. A case study in the state of Goias, Brazil," ,10.1101/2020.04.19.20071852, 2020.

[8] V. Jakkula, "Tutorial on support vector machine (svm)," School of EECS,Washington State University, 2006.

[9] Y. S. I. a. P. A. Mokhade, "Use of Linear Regression in Machine Learning for Ranking," IJSRD - International Journal for Scientific Research & Development, vol. 1, no. 5, 2013.

[10] S. Sperandei, "Understanding logistic regression analysis," Biochemia medica, vol. 24., pp. 12-8, 2014.

[11] W. a. Bruna, "Bayesian Linear Regression," 2019, 2019.

[12] D. J. S. ,. T. A. W. ,. J. D. C. ,. J. J. C. David R. Anderson, Quantitative Methods for Business 13 edition, Cengage Learning, 2015.

[13] C. S. M. S. Heumann, Introduction to Statistics and Data Analysis, Springer, 2016.

[14] W. K. H. O. Okhrin, Basic Elements of Computational Statistics, Springer, 2017.

[15] J. A. S. Betty R. Kirkwood, Essential Medical Statistics, Blackwell Science, 2003.

[16] WHO, "Confirmed Case , https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv ,"

[17] WHO, "Death, https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv".

[18] WHO, "Recoverd, https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv".

[19] D. Kirchman, "Calculating microbial growth rates from data on production and standing stocks," Marine Ecology-progress Series , 2002.

[20] J. &. L. Y. Lawson, "The Geometric Mean, Matrices, Metrics, and More," The American Mathematical Monthly, 2001.

[21] W. Y. a. M. Warshaure, "Arithmeyic and Gemometric Mean," Menemeni Matematik(descoviring mathematics), vol. 27, no. 2, pp. 17-22, 2002.

[22] T. &. D. Chai, "Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature," Geoscientific Model Development, vol. 7, 2014.

[23] K. M. Cort J. Willmott, "Advantages of the mean absolute error (MAE) overthe root mean square error (RMSE) in assessingaverage model performance," Climate Research Clim Re, vol. 30, pp. 79-82, 2005.

[24] P. &. B. C. &. S. L. Schober, "Correlation Coefficients: Appropriate Use and Interpretation," Anesthesia & Analgesia, 2018.

[25] A. D. Bland JM, Correlation, regression, and repeated, BMJ, 1994.