Journal of Cases on Information Technology
Volume 23 • Issue 4 • October-December 2021

Novel COVID-19 Mortality Rate Prediction (MRP) Model for India Using Regression Model With Optimized Hyperparameter

Dhamodharavadhani S., Periyar University, Salem, India
R. Rathipriya, Periyar University, Salem, India
https://orcid.org/0000-0002-3970-262X

ABSTRACT

The main objective of this study is to estimate the future COVID-19 mortality rate for India using COVID-19 mortality rate models from different countries. Here, the regression method with the optimal hyperparameter is used to build these models. In the literature, numerous mortality models for infectious diseases have been proposed, most of which predict future mortality by extending one or more disease-related attributes or parameters. But most of these models predict mortality rates from historical data. In this paper, the Gaussian process regression model with the optimal hyperparameter is used to develop the COVID-19 mortality rate prediction (MRP) model. Five different MRP models have been built for the U.S., Italy, Germany, Japan, and India. The results show that Germany has the lowest death rate in 2000 plus COVID-19 confirmed cases. Therefore, if India follows the strategy pursued by Germany, India will control the COVID-19 mortality rate even in the increase of confirmed cases.

KEYWORDS

COVID-19, Death, Hyperparameter, India, Mortality Prediction, Mortality Rate, Parameter Optimization, Regression Model, Tuning

1. INTRODUCTION

Currently, more than one million people around the world face the severe consequences of the outbreak of the novel Coronavirus 2019 (nCov). The first case of human infection by a nCov or Wuhan virus or 2019- nCov was reported in Wuhan, China[1]. The greatest challenge of this infectious disease is the human-to-human transition of nCov that would rise up the infected cases exponentially. On 30 January 2020, World Health Organization (WHO) issued a worldwide health emergency warning notice[2], designating that 2019-nCoV is of urgent global concern. The morbidity and mortality rates for the infection of 2019-nCoV are uncertain at the early stage (Sparrow 2020) especially for young ones and aged people. WHO has estimated the reproduction factor (R0) of nCov is 2.7. In order to control the wide and quick spread of the nCov, public health sectors took reliable preventative

DOI: 10.4018/JCIT.20211001.oa1

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

measures and imposed curfew or lockdown infested cities in China, United States, India, and other countries also. This is to limit the social distance among people and to avoid the transmission of this novel virus via humans to humans.

Over the past decade, machine learning techniques have gained momentum and have played an important role in many domains of research. Machine Learning (ML) is a subset of artificial intelligence that optimizes data through a series of mechanisms and provides novel insights or form of data to take timely active or preventive measures. In particular, it has a tremendous impact on data analysis and data science. It better understands the data and its processes, makes predictions about the future based on historical data / empirical data, and automatically classifies a group of data called classification. There are two types of learning techniques in ML techniques: unsupervised and unsupervised learning techniques.

Supervised learning techniques (regression, classification, and regression trees (CART) and nave Bayes) use labeled data to train input and output known mechanisms. Unsupervised learning techniques (Association Rules and Clustering.) use unlabeled data, inputting raw data directly to these methods without knowing the output of that data.

Machine learning techniques also can be used to develop standard mortality models. Deprez et al. (2017), used machine learning algorithms to fit and assess the mortality model. The regression approach used to detect the weaknesses of different mortality models by the authors. In Hainaut (2018), artificial neural networks (ANNs) used to find the latent factors of mortality and forecast them. Richman and Wüthrich (2018) extended the Lee-Carter model to multiple populations using neural networks.

Generally, Gaussian process models have been widely used in engineering based optimization applications (Razavi et al. 2012). In Raghavendra and Deka Raghavendra and Deka (2016), a combination of GPR and adaptive neuro-fuzzy inference system (ANFIS) used to forecast the ground-water level. In Roy and Datta (2018), an extensive comparative study was carried out between several surrogate models, comprising GPR, using simulation-optimization methodology with uncertainty pa- rameters. At the end, they had concluded that the GPR models and their ensemble were efficient methods with respect to prediction accuracy.

In this work, the Gaussian process regression model with the optimal hyperparameter is used to develop COVID-19 mortality models for five different countries (USA, Italy, Japan, Germany, and India). Also, evaluate the effectiveness of these models in the model evaluation and prediction of COVID-19 mortality rates for India. The purpose of this study is mortality prediction, which uses machine learning techniques, to clearly identify patterns that cannot be identified with the standard mortality model. The rest of this paper is plotted as follows. Section 2 describes the proposed method for predicting COVID-19 mortality for India with an early-stage time-series dataset. The results and discussion of the empirical study are presented in Section 3. Section 4 concludes this work with an extension of possible future work.

2. PROPOSED METHODOLOGY

Gaussian Process Regression (GPR) is a nonparametric kernel-based probabilistic model that has the ability to handle complex non-linear relations between response and predictor variables Ebden (2008). Gaussian process (GP) is a random process and is characterized as a set of random variables with a Gaussian joint multivariate distribution Hong and Zhou (2012); Rasmussen (2004); Ažman and Koci- jan (2011a); Huber (2014). A mean function and covariance function are used in GP. GP can achieve non-parametric regression function learning from noisy data and it have Gaussian distributions over the data Hong and Zhou (2012); Rasmussen and Williams (2006). The projected mean value is a linear combination by GP computation of the covariance function Hong and Zhou (2012); Saha et al. (2010). One of the applications of GPs is Gaussian process regression (GPR). GPR is deterministic and robust non-parametric Bayesian model that defines a priori distribution of

the likelihood over function space Rasmussen (2004); Ažman and Kocijan (2011a). It is the one of the most significant Bayesian machine learning methods that estimates the subsequent deterioration of non-linear regression by restricting the previous distribution to match the available training data Hong and Zhou (2012).

Usually, a GPR model is provided with training data and its performance is calculated by weighting targets in terms of error between training and test input Ažman and Kocijan (2011b). The output of the prediction is a Gaussian distribution of probability and is represented by its mean and variance. Vari- ance is the confidence factor for the output's expected mean value Hong and Zhou (2012); Ažman and Kocijan (2011a); Saha et al. (2010). Covariance function chosen by the user, but the hyperparameters can be learned from the training data using a gradient-based optimizer Hong and Zhou (2012).

2.1 Gaussian Process

Gaussian Process is a machine learning technique used to make predictions with uncertainty. It is defined as a finite collection of jointly Gaussian distributed random variables Hong and Zhou (2012). In regression problems, these random variables represent the values of a function f(x) at input x. It is represented as $\{f(x): x \in X\}$ which is defined by its mean function $\mu(x)$ and a covariance function $k(x, x')$ so that it can be written as:

$$f(\cdot) \sim GP(\mu(\cdot), k(\cdot,\cdot)) \tag{1}$$

Usually, the prior distribution over functions $f(\cdot)$ is a zero-mean Gaussian Process prior. The detailed explanation of the same found in Ebden (2008). In literature, this model is called a surrogate for the objective function. The surrogate is easier to optimize than the objective function. GP methods finds the next set of hyperparameters to evaluate on the actual objective function by selecting best hyperparameters that perform on this surrogate function.

2.1.1 Covariance Functions

This function is used to represents the covariance between pairs of random variables in GPR. It can be written as in equation (2):

$$K_{ij} = k(x_i; x_j) = \alpha \exp\left\{\frac{\left\|x_i - x_j\right\|^2}{2\sigma_1^2}\right\} \tag{2}$$

Here, hyperparameters are:

σ_1 = *Characteristic lengthscale*
α = *Signal variance*

2.1.2 Merits of GPR

- It has high adaptability and accurate prediction for processing small data set and also for high-dimensional data Hong and Zhou (2012); Ažman and Kocijan (2011a); Wang et al. (2008).
- It can learn noisy data using non-parametric regression function and, avoiding simple parametric assumptions Huber (2013).
- It can handle small dataset as well as data with uncertainty very effectively and efficiently.

2.2 Hyperparameter Optimization

In regression models, the objective of hyperparameter optimization is to find the parameters of a given algorithm that return the best performance on a validation set while training and testing the model, Hong and Zhou (2012). It is represented as (equation 3):

$$H^* = \arg\min_{o \in O} f(o) \tag{3}$$

- f(x) represents an objective score to minimize root mean squared error (RMSE) evaluated on the validation set;
- x* is the set of hyperparameters which yields the lowest score of RMSE; and
- x is any value in the problem domain X.

Even though hyperparameter optimization is extremely expensive in terms of the computational time, it yields good prediction accuracy than traditional regression models.

GPR model-based hyperparameter optimization has following steps:

Step 1: Initialize hyperparameters for GPR model based on the problem.
Step 2: An objective function of GPR model which takes in these hyperparameters and outputs a RMSE score that has minimal value.
Step 3: Define surrogate model of the objective function.
Step 4: Specify the selection criteria for evaluating hyperparameters which have to choose next from the surrogate model.
Step 5: Maintain the history of (score, hyperparameter) pairs used by the GPR algorithm to update the surrogate model.
Step 6: Repeat steps 2–5 until maximum iterations or time is reached.

2.3 Polynomial Regression Model

Polynomial regression is a type of regression model in which the maximum degree of the predictor variable is more than 1. In this technique, the best fit line is in the curve shape. It can be used to approximate a complex nonlinear relationship Seyedzadeh et al. (2018).

The k th order polynomial model in one variable is given by equation (4):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdot + \beta_k x^k + \varepsilon \tag{4}$$

where:

$$x_j = x_j,\ j = 1,2,..,k$$

then the model in k explanatory variables $x_1, x_2, , . . ., x_k$ is multiple linear regression.

If a polynomial regression model is of $y = X\beta + \epsilon$ then the techniques for fitting linear regression model can be used for fitting the polynomial regression model.

2.4 RMSE

RMSE is a polynomial scoring rule which also measures the error average size Hong and Zhou (2012), Dhamodharavadhani and Rathipriya (2020a), Dhamodharavadhani and Rathipriya (2020b) This is the square root of the square differences measured between prediction and actual observation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - A_i\right)^2} \tag{5}$$

It represents as:

n= number of samples P=Predicted values A=Actual values

In this study, the GPR method is explored as a potential regression model for dealing with non-linear variability in predicting COVID-19 mortality for India. For comparison purposes, a polynomial regression method is employed. The final results demonstrated that the GPR method, although time consuming, is the most efficient in terms of the Root Mean Square Error (RMSE), which is the main performance measurement of regression models. Figure 1 clearly depicts the flow of the proposed method.

The best and worst-case scenarios for COVID-19 spread across the globe are taken for study to model the COVID-19 mortality rate for India. For that purpose, data on confirmed cases and death cases of five countries (such as the United States, Italy, Japan, Germany, and India) are used to create a model. This information is taken from the website www.kaggle.com.Japan and Germany are being cited as the best-case scenario for the COVID-19 outbreak and its deaths, while the US and Italy are the worst-case scenarios for the COVID-19 outbreak and its death. China was not considered in this study because their COVID-19 outbreak was restricted. Models for the mortality rate of COVID-19 developed using the Polynomial Regression Model and the GPR Regression Model for these countries. RMSE is used as a key quality performance indicator to select the best model. Finally, predict the mortality rate for India using the best mortality rate model.

3. RESULTS AND DISCUSSIONS

An experiment is carried out to evaluate the efficiency of the proposed methodology for Morality Rate Prediction of the COVID-19 pandemic in India. Empirical data is used, which is extracted from the Kaggle Website. Since 21 Jan2020, the data is updated daily basis, of the increment of the number of infested people in COVID-19 infected countries across the world (Table 1). The experimentation platform is an i7-CPU running 64-bits OS of MS Windows 10. In Figure 2, X-axis represents the date and Y-axis represents the number of COVID-19 cases. It clearly shows the number of active cases, the number of death cases, and the number of recovered cases against the date. List of top 10 countries having higher rate of COVID-19 confirmed cases shown in Figure 3. Figure 4 represents number of COVID-19 death cases for countries such as US, Italy, Japan, Germany, and India. Here, X-axis indicates the date and Y-axis shows the number of COVID-19 cases (Table 2).

Table 3 shows the Kernel Parameters such as sigmaM, sigmaF, and sigma of the best objective function for GPR models of various countries. Squared Exponential used as kernel function for this GPR model. Each country has different hyperparameter tuning value which reflects that each country has a different trend in the COVID-19 spread and have the rate of increased confirmed cases also different. The regression loss for predicted value using GPR model for five countries are also given in Table 4. RMSE for the same are also tabulated. Table 5 shows the RMSE value for Polynomial Regression (PR) model of five countries. Figure 5 shows the comparison of RMSE value for Polynomial Regression and GPR model. In this figure, X-axis represents the countries and Y-axis shows the RMSE value.

It indicates that GPR model have very low RMSE than Polynomial Regression model for five countries. Hence therefore, GPR model is used to predict COVID-19 morality rate for India.

Figure 1. Proposed Methodology for COVID-19 MRP for India

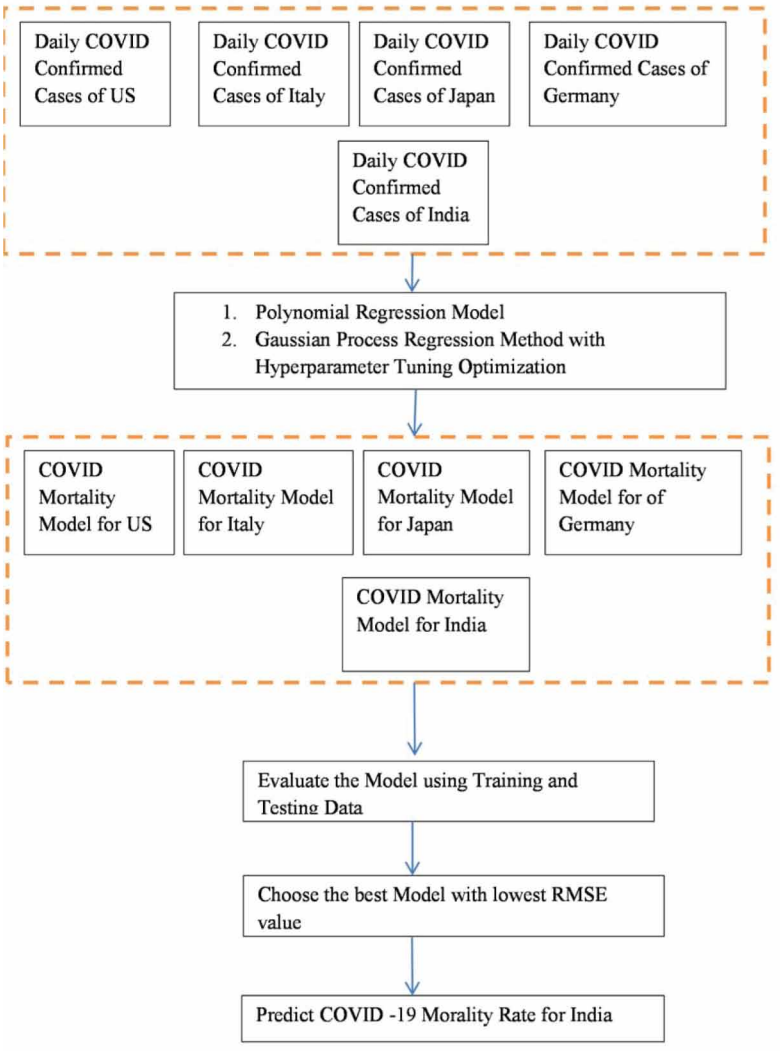


Table 1. Hyperparamters of GPR Model for different countries

Kernel Parameters	US	Japan	Germany	Italy	India
SigmaM	256876.3	589.8239	22374.43	79290.37	2667.455
SigmaF	7008.68	20.46376	332.0723	10241.42	50.8837
Sigma	37.07613	1.110905	4.428102	87.57416	1.654951

Table 6 represents the number of predicted COVID-19 death cases for countries like US, Italy, Japan, Germany, and India using GPR and PR model respectively. Figure 6 illustrates the ranking of COVID- 19 mortality rate model of five countries. Here, X-axis indicates RMSE value and Y-axis shows the country's name. It is clearly known that India has very low RMSE value because the magnitude of the confirmed cases is low. At the same time, mean mortality rate for India is the second

Table 2. Loss and Error Value for GPR Model

	US	Japan	Germany	Italy	India
Regression Loss	179.1848	1.878231	3.009694	1811.484	0.634793
RMSE	13.38599	1.370486	1.734847	42.56153	0.796739

Table 3. RMSE value for Polynomial Regression Model

US	Japan	Germany	Italy	India
137.567	3.227635	57.21517	439.4404	1.445828

Table 4. Predicted COVID-19 Death cases using GPR model

Number of Confirmed Cases	US	Japan	Germany	Italy	India
1500	29	56	3	91	57
2000	35	53	4	119	73
2500	35	53	4	119	73
3000	41	37	5	147	86
3500	47	29	6	175	96
4000	53	28	7	204	101
4500	59	28	8	234	102
5000	65	28	10	264	99

Figure 2. Breakdown of COVID-19 Confirmed Cases

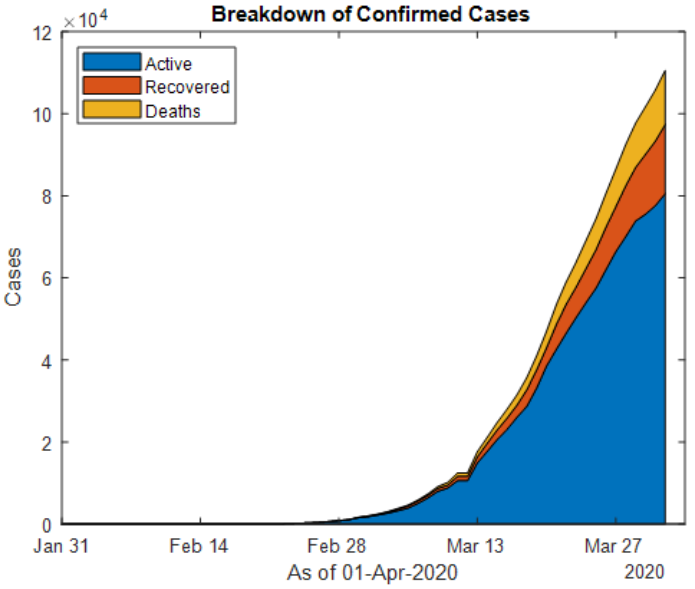
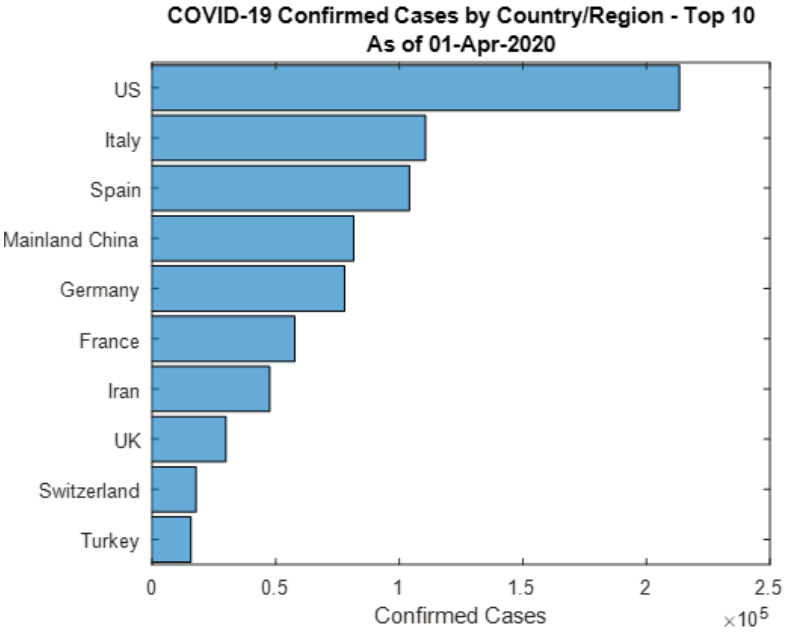


Table 5. Predicted COVID-19 Death cases using PR model

Number of Confirmed Cases	US	Japan	Germany	Italy	India
2000	20	76	5	46	65
2500	20	76	5	46	65
3000	30	92	9	101	78
3500	39	108	14	155	91
4000	49	123	18	209	104
4500	58	139	22	264	117
5000	68	154	27	318	131

Figure 3. Top 10 Countries-COVID Confirmed Cases



highest among the five countries taken for study. Therefore, Japan is the second top RMSE value for the MRP model, which is selected as the best model. While the number of COVID-19 confirmed cases increases, number of death cases for Germany is very low when compared to other countries in those tables. But, at the same time, it shows the increase pattern of death cases. For Japan, the number of death cases is not increased after 4000 confirmed cases. Moreover, RMSE value for Japan's COVID-19 Mortality model also low than Germany. In Figure 7, X-axis show day of COVID-19 since first day and Y-axis indicates the number of predicted COVID-19 death case. Figure 7 shows the number of predicted death cases for five countries. Table 6 shows the mortality rate prediction for US, Japan, Germany, Italy, and India respectively. Figure 8 represents the ranking of five countries based on mean mortality rate. Here, X-axis indicates mean mortality rate value and Y-axis shows the country's name. Based on the mean mortality rate, Germany's model is the best model for MRP.

Figure 4. Number of COVID-19 Death cases for five countries

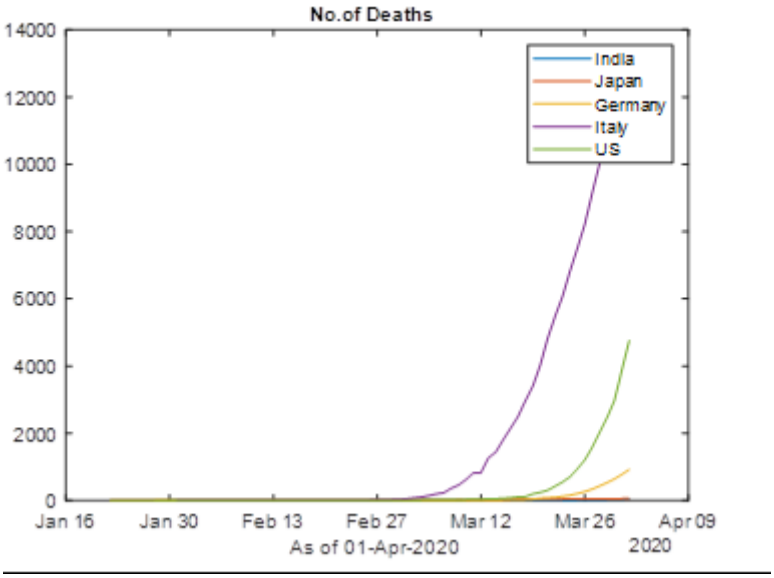
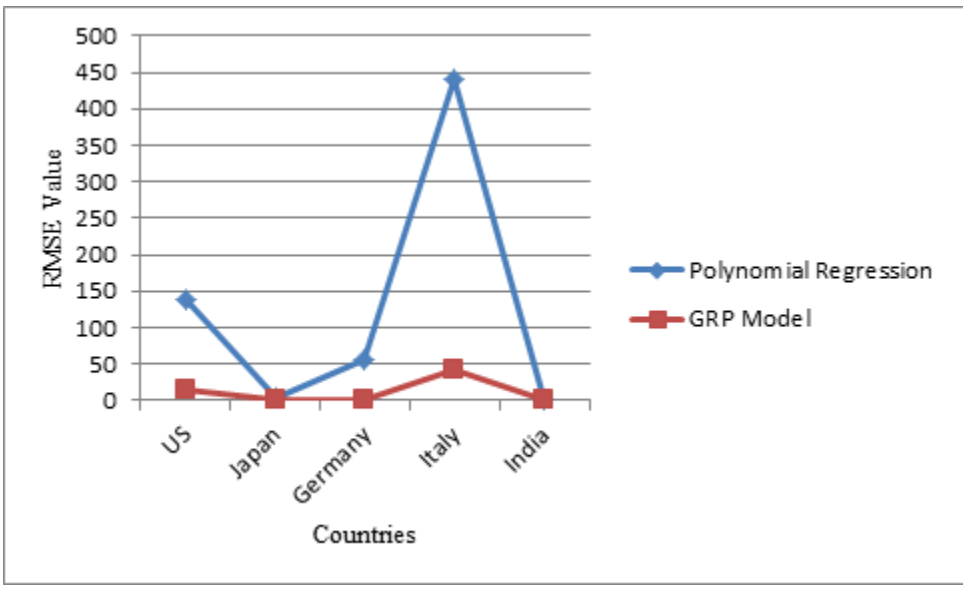


Figure 5. Comparison of RMSE value



3.1 Suggestions Based on This Study

The social distancing; COVID-19 awareness and best self-hygienic practices are very important factors to constraint COVID-19 deaths in that country. Therefore, Japan's Mortality model is selected as the best model for predicting COVID-19 death cases. If India also adopts Japan's and Germany's strategies along with national lockdown, India's COVID-19 death cases can be reduced in the future.

Figure 6. RMSE Value for Five Countries Mortality Rate Model

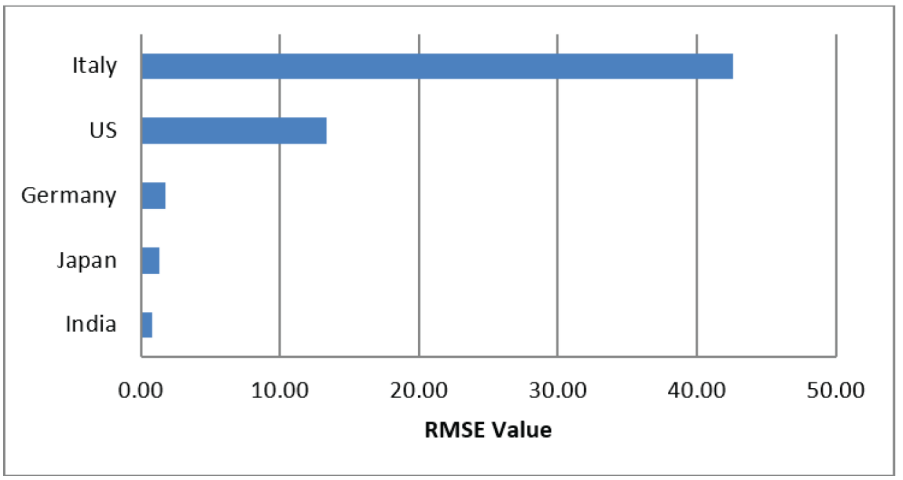
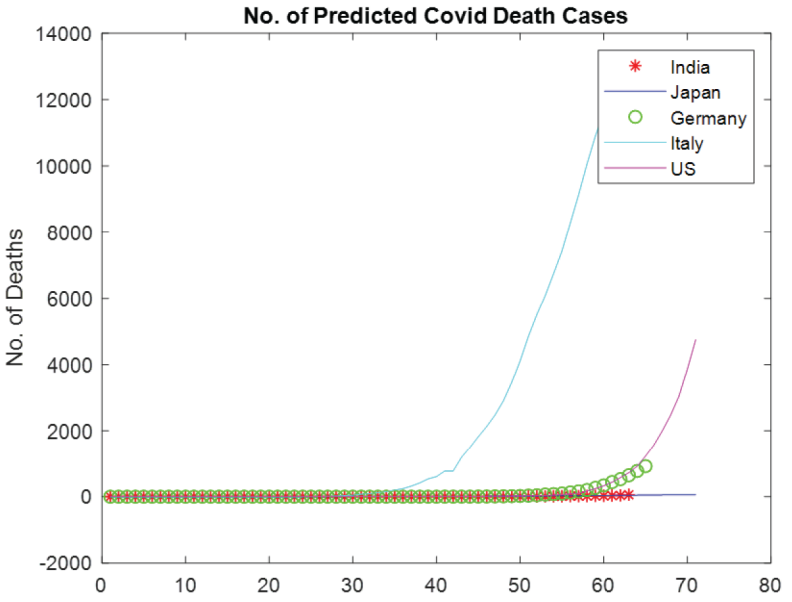


Figure 7. Predicted COVID-19 Death Cases using GPR model



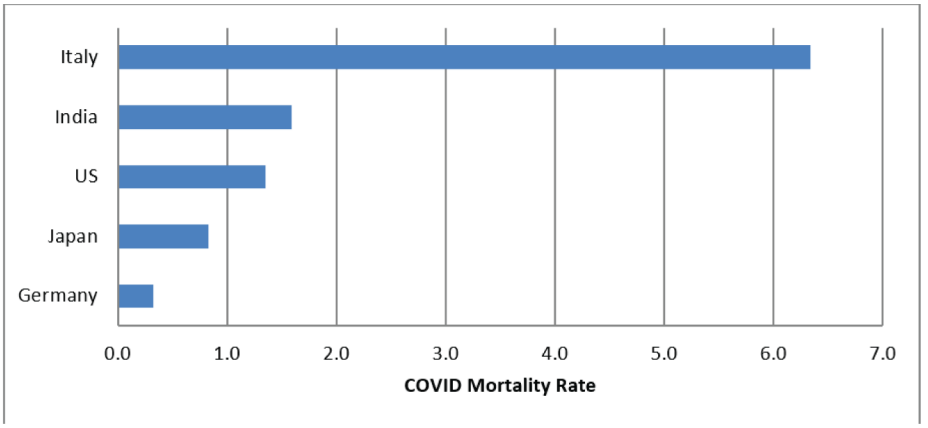
4. CONCLUSION

The present study has examined the performance of Gaussian process regression with hyperparameter optimization and Polynomial Regression for COVID-19 Confirmed cases dataset. It contains a number of confirmed cases and death cases for the period 22, January 2020 to 01, April 2020. COVID-19 Mortality model for US, Italy, Japan, Germany, and India had been build using GPR and PR model. Their performance had been analyzed using RMSE value. The evaluation results indicate that the GPR

Table 6. COVID-19 Mortality Rate Prediction for Five Countries

No. of Confirmed Cases	US	Japan	Germany	Italy	India
2000	1.5	2.8	0.2	4.6	2.9
2500	1.4	2.1	0.2	4.8	2.9
2500	1.4	2.1	0.2	4.8	2.9
3000	1.4	1.2	0.2	4.9	2.9
3500	1.3	0.8	0.2	5.0	2.7
4000	1.3	0.7	0.2	5.1	2.5
4500	1.3	0.6	0.2	5.2	2.3
5000	1.3	0.6	0.2	5.3	2.0
10000	1.3	0.3	0.3	6.0	0.5
20000	1.3	0.1	0.3	7.0	0.2
30000	1.3	0.1	0.5	7.7	0.2
40000	1.3	0.1	0.6	8.4	0.1
50000	1.3	0.1	0.7	8.9	0.1
100000	1.5	0.0	0.9	11.3	0.0

Figure 8. Ranking of Countries based on Mean Mortality Rate



method based Mortality model for India has low RMSE but it has higher mortality rate. Therefore, morality rate based MRP model was selected to contained COVID-19 death rate in India. Hence, it is found that the COVID-19 MRP model of Germany is the best model for controlling the COVID-19 mortality rate in India.

ACKNOWLEDGMENT

The first author acknowledges the UGC- Special Assistance Programme (SAP) for the financial support to her research under the UGC-SAP at the level of DRS-II (Ref.No.F.5-6/2018/DRS-II (SAP-II)), 26 July 2018 in the Department of Computer Science.

REFERENCES

Ažman, K., & Kocijan, J. (2011). Dynamical systems identification using Gaussian process models with incorporated local models. *Engineering Applications of Artificial Intelligence*, *24*(2), 398–408. doi:10.1016/j.engappai.2010.10.010

Deprez, P., Shevchenko, P. V., & Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. *European Actuarial Journal*, *7*(2), 337–352. doi:10.1007/s13385-017-0152-4

Dhamodharavadhani, S., & Rathipriya, R. (2020a). Enhanced Logistic Regression (ELR) Model for Big Data. Handbook of Research on Big Data Clustering and Machine Learning, 152–176.

Dhamodharavadhani, S., & Rathipriya, R. (2020b). Variable Selection Method for Regression Mod- els Using Computational Intelligence Techniques. Handbook of Research on Machine and Deep Learning Applications for Cyber Security, 416–436.

Ebden, M. (2008). Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30. PMID:18084059

Hainaut, D. (2018). A neural-network analyzer for mortality forecast. *ASTIN Bulletin*, *48*(02), 481–508. doi:10.1017/asb.2017.45

Hong, S., & Zhou, Z. (2012). *Application of Gaussian process regression for bearing degradation assessment. Information Science and Service Science and Data Mining*.

Huber, M. (2014). Recursive Gaussian process: On-line regression and learning. *Pattern Recognition Letters*, *45*, 85–91. doi:10.1016/j.patrec.2014.03.004

Huber, M. F. (2013). Recursive Gaussian process regression. *IEEE Int. Conf. Acoust. Speech Signal Process*.

Raghavendra, N. S., & Deka, P. C. (2016). Multistep ahead groundwater level time-series forecasting using Gaussian Process Regression and ANFIS. In Advances in Intelligent Systems and Computing (vol. 396). Springer. doi:10.1007/978-81-322-2653-6_19

Rasmussen & Williams. (2006). *Gaussian processes for machine learning*. MIT Press.

Rasmussen, C. E. (2004). Gaussian processes in machine learning. *Adv. Lect. Mach. Learn*, *14*(2).

Razavi, S., Tolson, B. A., & Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, *48*(7). Advance online publication. doi:10.1029/2011WR011527

Richman, R., & Wüthrich, M. V. (2018). *A Neural Network Extension of the Lee-Carter Model to Multiple Populations*. SSRN. doi:10.2139/ssrn.3270877

Roy, D. K., & Datta, B. (2018). Trained meta-models and evolutionary algorithm based multi- objective management of coastal aquifers under parameter uncertainty. *Journal of Hydroinformatics*, *20*(6), 1247–1267. doi:10.2166/hydro.2018.087

Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: A review. *Visualization in Engineering*, *6*(1), 5–5. doi:10.1186/s40327-018-0064-7

Sparrow, A. (2020). "How China's Coronavirus Is Spreading—and How to Stop It". *Foreign Policy*.

Wang, J. M., Fleet, D. J., & Hertzmann, A. (2008). Gaussian Process Dynamical Models for Human Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(2), 283–298. doi:10.1109/TPAMI.2007.1167 PMID:18084059

ENDNOTES

1 "WHO | Novel Coronavirus – China". WHO. Archived from the original on 23 January 2020. Retrieved 1 February 2020.

2 Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)". World Health Organization (WHO). 30 January 2020. Archived from the original on 31 January 2020. Retrieved 30 January 2020.