Applied Intelligence (2021) 51:2703–2713
https://doi.org/10.1007/s10489-020-01942-7

Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model

Mohd Saqib[1]

Accepted: 11 September 2020 / Published online: 23 October 2020

Abstract
In 2020, Coronavirus Disease 2019 (COVID-19), caused by the SARS-CoV-2 (Severe Acute Respiratory Syndrome Corona Virus 2) Coronavirus, unforeseen pandemic put humanity at big risk and health professionals are facing several kinds of problem due to rapid growth of confirmed cases. That is why some prediction methods are required to estimate the magnitude of infected cases and masses of studies on distinct methods of forecasting are represented so far. In this study, we proposed a hybrid machine learning model that is not only predicted with good accuracy but also takes care of uncertainty of predictions. The model is formulated using Bayesian Ridge Regression hybridized with an n-degree Polynomial and uses probabilistic distribution to estimate the value of the dependent variable instead of using traditional methods. This is a completely mathematical model in which we have successfully incorporated with prior knowledge and posterior distribution enables us to incorporate more upcoming data without storing previous data. Also, L^2 (Ridge) Regularization is used to overcome the problem of overfitting. To justify our results, we have presented case studies of three countries, −the United States, Italy, and Spain. In each of the cases, we fitted the model and estimate the number of possible causes for the upcoming weeks. Our forecast in this study is based on the public datasets provided by John Hopkins University available until 11th May 2020. We are concluding with further evolution and scope of the proposed model.

Keywords COVID-19 pandemic · Bayesian ridge regression · Prediction · Mathematical modeling

1 Introduction

In late December 2019, a group of patients was come up with an unknown Etiology to the hospitals having symptoms of pneumonia. Later on, the first case of novel coronavirus was reported in the city of Wuhan in Hubei province in Central China [1]. After taking a basic understanding of the virus, medical experts have given a name as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the name of the disease caused by this virus is coronavirus disease 2019 (COVID-19) [2]. The cases of COVID-19 pandemic are growing rapidly. Till 30th April 2020, we have 3,251,587 confirmed and 229,832 death cases throughout the world due to this hazardous pandemic, COVID-19.

In India, the first laboratory-confirmed case of COVID-19 was reported from Kerala on 30th January 2020 and as of 30th April 2020, a total of 33,931 cases and 943 deaths were reported in India [3]. To tackle this ongoing pandemic and such events in the future where the lives of millions of people are at high risk, we need a strong health care system and technology that will be the means of making a way to a panacea. Whenever such pandemic spread in a country or province it has some patterns and various mathematical models can be proposed to forecast using such technologies and mathematical theories. For example, in [4], the authors proposed a model for Malaria transmission dynamics of the anopheles mosquito and in [5] a Bifurcation analysis for malaria transmission has been developed. As we are also aware of the menacing of HIV/TB and study [6] presented the mathematical analysis of the transmission dynamics of the same. According to [7], Due to being class of β-Coronavirus, it has a spreading capability among hosts (Primary to secondary source) and that is why the magnitude of infected cases growing non-linearly. Non-linearity of any pandemic can be detected in several ways e.g. in [8] a laplacian based decomposition is used to solve the non-linear parameters in a Pine Witt disease. Similarly, in [9], a fractional version of SIRS (Susceptible - Infectious - Recovered - Susceptible) model has been developed to help,

✉ Mohd Saqib
msaqib.cs@gmail.com

[1] Mathematic and Computing Department, Indian Institute of Technology (Indian School of Mines), Dhanbad, Jharkhand, India

to control the syncytial virus in infants. Also, in [10], the author has used Generalized Additive Models (GAMs) to predict dengue outbreaks based on disease surveillance, meteorological and socio-economic data.

Despite several research works and their documentation, there are huge opportunities for the utilization of AI, Machine Learning, and Data Science in this field, due to the novelty of the root cause. For example, in [11]. This article author has a comprehensive discussion regarding AI applications, constraints, and pitfalls during the COVID-19 pandemic. So, there must be some prediction methods that are required to estimate the magnitude of infected cases, and masses of studies on distinct methods of forecasting are represented so far [12]. In [12, 13], authors estimate the possible number of infected cases in India using long short-term memory (LSTM). Same as in [14], the study represented virus progression and forecast using the same algorithm for Canada and compare with the United States (US) and Italy. In [15], Sujatha performed linear regression (LR), Multilayer perceptron (MLP), and Vector autoregression model (VARM) for expectation on the COVID-19 kaggle information to anticipate the epidemiological pattern of the disease and rate of COVID-2019 cases in India. In [16] author proposed machine learning models (XGBoost and Multi-Output Regressor) to predict confirmed cases over the coming 24 days in every province of South Korea with 82.4% accuracy. As we have already discussed, a study in [10], proposed to control the syncytial virus in infants, same as for China, a modified SEIR and AI prediction of the trend of the epidemic of COVID-19 has been proposed in this study [17]. Different research also takes place on the cases of India but using different methods, and autoregression integrated moving average model (ARIMA) and Richard's model [18]. Moreover predictions, some mathematical models have also estimated the effects of lockdown and social-distancing in India in a practical scenario [19] but all these studies represented so far are based on inadequate data at the initial stage without any measurement of uncertainty. These models are developed with good accuracy but as well as the data become available, those entire algorithms will not be able to survive without a few evaluations due to the dynamic nature of pandemic escalation of the COVID-19.

So, a distribution based learning model will be more promising rather than doing point estimation. Bayesian Learning is a very well-known method of making any prediction based on our prior knowledge [20]. Many studies have been already used the Bayesian approach for prediction for many pandemics and clinical forecasting like in [21] authors have been estimated the probability of demonstrating vaccine efficacy in the declining Ebola epidemic using the Bayesian modeling approach. In this [22] chapter, the author focuses on the various utility of Bayesian Prediction and it is not only useful, but simple, exact, and coherent, and hence beautiful. Also, the study [23] illustrated a Bayesian analysis for emerging infectious diseases. Same as in [24], paper presented a Bayesian scheme for real-time estimation of the probability distribution of the effective reproduction number of the epidemic potential of emerging infectious diseases and show how to use such inferences to formulate significance tests on future epidemiological observations. Besides, a study also proposed a system, able to provide early, quantitative predictions of SARS epidemic events using a Bayesian dynamic model for influenza surveillance demonstrated [25]. So, this was the motivation behind the proposed study, the prediction of infected cases by COVID-19 which is also a SARS family virus can be formulated using Bayesian learning as a study [25] already represented for influenza surveillance. In the proposed study we are formulating Bayesian Learning Regression with a polynomial of n-degree. Furthermore, one issue occurs when working with time-series data (as COVID-19 confirmed cases) is over-fitting particularly when estimating models with large numbers of parameters over relatively short periods and the solution to the over-fitting problem, is to take a Bayesian approach (using Ridge Regularization) which allows us to impose certain priors on depended variables [26]. Another big reason we often prefer to use Bayesian methods is that it allows us to incorporate uncertainty in our parameter estimates which are particularly useful when forecasting [26].

The manuscript is organized as follows. "**Method and Model**" explains the methodology used to construct the model and various terminology used in the study. " **Significance of Proposed Model in COVID-19 Outbreak**" describes the important advantages of such a hybrid model and also discussed our novelty of the work in the COVID-19 outbreak. After that three case studies in "**Case Studies**" also presented to justify our results and fruition of the model. In the last, we discussed our results, comparison with other developed models, and finding in the section "**Results and Discussion**" followed by the conclusion in "**Conclusion**".

2 Method and model

2.1 Datasets

The datasets collected from Johns Hopkins University are used in the studies [27]. The datasets accessed on 11 May 2020. It provides several fatalities and registered patients by the end of each day. The dataset is available in the time series format with date, month, and year so that the temporal components are not neglected. A wavelet transformation [28] is applied to preserve the time-frequency components and it also mitigates the random noise in the dataset. This dataset consists of six columns (Table 1).

The only pre-processing was required to transform the dataset. The observations recorded every day and for each

Table 1 Dataset Description

S.R.	Column Name	Data Type	Data Description
1	ID	INT32	Unique ID for each day
2	Province_State	String	Name of state
3	Country_Region	String	Name of country
4	Date	Date	Date of each day starts from 22, Jan 2020
5	Confirmed Cases	INT64	Total No. of cases found till the date
6	Fatalities	INT64	Total No. of deaths occurred till the date

day a new column added. The datasets are divided into two parts training (80%) and testing (20%) datasets.

2.2 Model formulation

One of the very basic approaches to make a prediction is another version of linear regression is Polynomial regression in which the relationship between independent and dependent variables is an n-degree polynomial. Mathematically, it can represent as follows:

$$f(X) = \beta_0 + \beta_1 x_1^1 + \beta_2 x_2^2 + \ldots + \beta_n \; x_n^n + \epsilon \tag{1}$$

Or,

$$f(X) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i^i + \epsilon \tag{2}$$

Where $\beta_i$is the coefficient and ϵ the measurement error which is

$$\epsilon \sim N\left(0, \sigma^2\right) \tag{3}$$

$f(X)$ is our polynomial model and to develop a good model we need to tuning, β_i So that following loss function with L^2 Regularization (Ridge Regularization) will be as minimum as possible

$$\beta = L(y_i, x_i) = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 + \lambda \sum_{i=1}^{n} \beta_i^2 \tag{4}$$

Where, the first part of the Eq. 4 is the residual sum of squares (RSS), the difference between actual value (y_i) and predicted value ($f(x_i)$) of the i^{th} observation. λ is the regularization term, deciding how much regularize the β_i.

Now, for the best fitting our aim to minimize the β by tuning coefficients, β_i. According to [29], the Maximum Likelihood Estimate of β which reduces the $L(y_i, x_i)$ is

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{x^T x}\right)^{-1} \boldsymbol{x^T y} \tag{5}$$

Now, instead of a vector of coefficients, we have a single value $\widehat{\beta}$, in $\mathbb{R}^{p+1}$ [30]. Here Bayesian Regression (BR) comes into the picture. In the BR, instead of predicting value mentioned as above, it used probabilistic distribution to estimate the value of $y_i$and its follow the following syntax

$$y_i \sim N\left(\beta^T X, \sigma^2\right) \tag{6}$$

So,

$$\boldsymbol{p(y \mid X, \beta, \sigma^2)} \propto \frac{\mathbf{1}}{\sqrt[n/2]{\boldsymbol{\sigma^2}}} \boldsymbol{e}^{-\frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)} \tag{7}$$

From conjugate prior distribution [20],

$$\begin{aligned} &(y - X\beta)^T (y - X\beta) \\ &= \left(y - X\widehat{\beta}\right)^T \left(y - X\widehat{\beta}\right) - \left(\beta - \widehat{\beta}\right)^T \left(X^T X\right) \left(\beta - \widehat{\beta}\right) \end{aligned} \tag{8}$$

The Eq. 7 is re-written as

$$\begin{aligned} p(y \mid X, \beta, \sigma^2) &\propto \frac{1}{\sqrt[n/2]{\sigma^2}} exp\left(-\frac{vs^2}{2\sigma^2}\right) \left(\sigma^2\right)^{-\frac{n-v}{2}} \\ &exp\left(-\frac{1}{2\sigma^2} \left(\beta - \widehat{\beta}\right)^T \left(X^T X\right) \left(\beta - \widehat{\beta}\right)\right) \end{aligned} \tag{9}$$

Where, $vs^2 = \left(y - X\widehat{\beta}\right)^T \left(y - X\widehat{\beta}\right)$ and $v = n - k$, n is the number of observations, and k is the number of coefficients in vector β. This suggests a form for the prior distribution is

$$p\left(\beta, \sigma^2\right) = p\left(\sigma^2\right) p\left(\beta \mid \sigma^2\right) \tag{10}$$

After the formulation of the prior distribution, now we need to generate posterior distribution, which can be formulated as follow (from Eqs. 7, 9 and 10),

$$\begin{aligned} &p\left(\beta, \sigma^2 \mid y, X\right) \propto p\left(y \mid X, \beta, \sigma^2\right) p\left(\beta \mid \sigma^2\right) p\left(\sigma^2\right) \\ &p\left(\beta, \sigma^2 \mid y, X\right) \propto \frac{1}{\sqrt[n/2]{\sigma^2}} e^{-\frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)} \\ &\left(\sigma^2\right)^{-\frac{n-v}{2}} e^{\left(-\frac{1}{2\sigma^2}(\beta - \mu_0)^T \Lambda_0 (\beta - \mu_0)\right)} \left(\sigma^2\right)^{-(a_0 - 1)} e^{\left(\frac{b_0}{\sigma^2}\right)} \end{aligned} \tag{11}$$

Where Λ_0 is ridge regression [31] used to overcome the problem of multicollinearity normally occurring when the model has large numbers of parameter and it is equal to

$$\Lambda_0 = cI \tag{12}$$

In Equation 12, $c = \sum_{i=1}^{n} \beta_i$ and I is an identity matrix of $n \times n$. Now, the posterior mean (μ_n) can be represented in the term of $\widehat{\beta}$ and prior mean μ_0 and for the Bayesian learning other can be upgraded as follows

$$\begin{aligned} \mu_n &= \left(X^T X + \Lambda_0\right)^{-1}\left(X^T X\widehat{\beta} + \Lambda_0\mu_0\right) \\ \Lambda_n &= \left(X^T X + \Lambda_0\right) \\ a_n &= \left(a_0 + \frac{n}{2}\right) \\ b_n &= \left(b_0 + \frac{1}{2}\left(y^T y + \mu_0^T \Lambda_0 \mu_0 + \mu_n^T \Lambda_n \mu_n\right)\right) \end{aligned} \tag{13}$$

Now we are ready to estimate the probability of y on given conditions (m) using Bayes Theorem as

$$p(y \mid m) = \frac{p(\beta, \sigma|m)p(y|X, \beta, \sigma, m)}{p(\beta, \sigma|y, X, m)} \tag{14}$$

Where m is the marginal likelihood and prior density, here, m is $p(y|X, \beta, \sigma)$ (See Fig. 1).

2.3 Parameter setting

There are many parameters used in the proposed model (Table 2) and a Fit-and-Score method implemented to optimize. It also implements Predict, Predict_proba, Decision_function, Transforms, and Inverse_transform if they are implemented in the estimator used. The parameters of the estimator used to apply these methods are optimized by cross-validated search over parameter settings [32]. The number of parameter settings that are tried is given by n_iter (≈100 in the proposed model). We initialize the parameters with default values and obtain the best-fitted parameters as given in the following Tables (Tables 3 and 4). The optimization of hyperparameters take place by implementing proposed model in Python.3.6 using scikit-learn [32] and used Spyder, a publically available software, a GUI to debug the code. The piece of code available as follows.

Table 2 Model Parameters

Parameter & Description
n_iter – int, optional It represents the maximum number of iterations. The default value is 300 but the user-defined value must be greater than or equal to 1.
tol – float, optional, default = 1.e-3 It represents the precision of the solution and will stop the algorithm if w has converged.
alpha_1 – float, optional, default = 1.e-6 It is the 1st hyperparameter which is a shape parameter for the Gamma distribution prior over the alpha parameter.
alpha_2 – float, optional, default = 1.e-6 It is the 2nd hyperparameter which is an inverse scale parameter for the Gamma distribution prior over the alpha parameter.
lambda_1 – float, optional, default = 1.e-6 It is the 1st hyperparameter which is a shape parameter for the Gamma distribution prior over the lambda parameter.
lambda_2 – float, optional, default = 1.e-6 It is the 2nd hyperparameter which is an inverse scale parameter for the Gamma distribution prior over the lambda parameter.

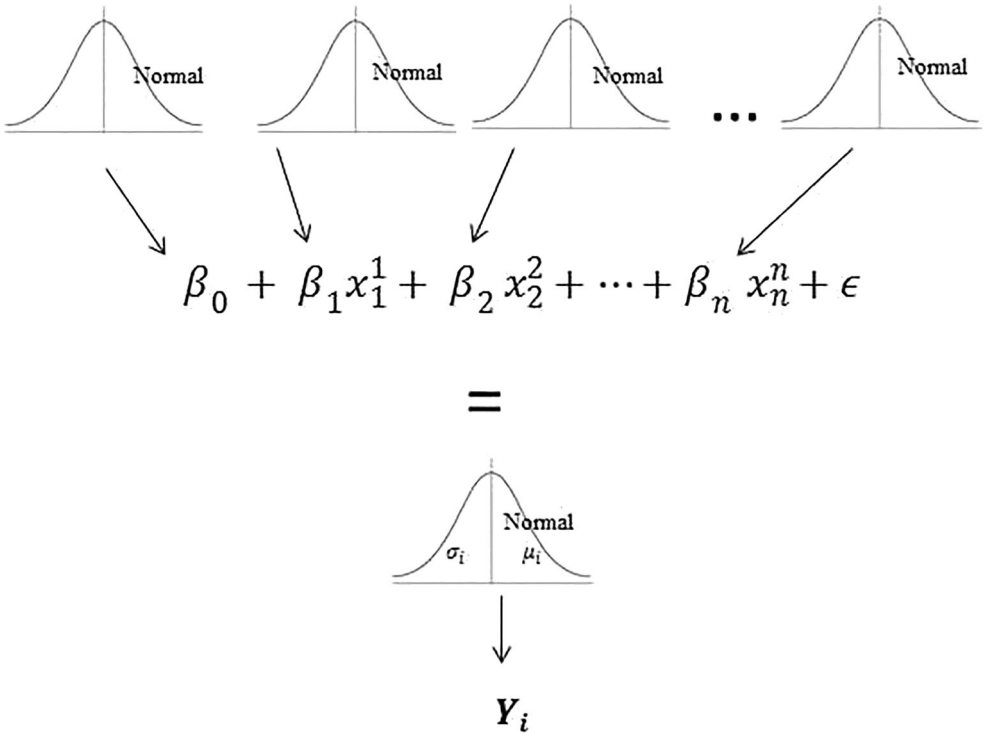

Fig. 1 PBRR Demonstration

Table 3 Initial Parameters Values of the Model

Country	Initial Parameter Values				
	tol	alpha_1	alpha_2	lambda_1	lambda_2
Italy U.S. Span	1.00E-04	1.00E-07	1.00E-07	1.00E-06	1.00E-06

tol = [1e-4, 1e-3, 1e-2]
alpha_1 = [1e-7, 1e-6, 1e-5, 1e-4]
alpha_2 = [1e-7, 1e-6, 1e-5, 1e-4]
lambda_1 = [1e-7, 1e-6, 1e-5, 1e-4]
lambda_2 = [1e-7, 1e-6, 1e-5, 1e-4]

bayesian_grid = {'tol': tol, 'alpha_1': alpha_1, 'alpha_2': alpha_2, 'lambda_1': lambda_1, 'lambda_2': lambda_2}.

bayesian = BayesianRidge(fit_intercept = False, normalize = True).

bayesian_search = RandomizedSearchCV(bayesian, bayesian_grid, scoring = 'neg_mean_squared_error', cv = 3, return_train_score = True, n_jobs = −1, n_iter = 40, verbose = 1).

bayesian_search.fit(poly_X_train_confirmed, y_train_confirmed).

bayesian_search.best_params_.

2.4 Advantages and novelty of the work

In the proposed model we have developed concepts of Bayesian inference that differ fundamentally from the traditional approach. This is completely mathematical methods in which we have successfully incorporated with prior knowledge. Instead of making predictions only, it discovers full probability distribution of the problem-domain even on a small dataset which also encounters the features of the confidence interval, risk aversity, etc. [33]. Moreover, posterior distribution makes the model to incorporate more upcoming data without storing previous data. In the current situation of the pandemic, data are not enough to make any prediction without any measurement of uncertainty. In the introduction section, we have seen many studies for COVID-19 progression with good accuracy but as well as data become available, those entire algorithms will not able to survive without a few evaluations. It will happen because of the dynamic nature of pandemic escalation. For example, if we consider our traditional regression methods (Eq. 1)

$$f(X) = \beta_0 + \beta_1 x_1^1 + \beta_2 x_2^2 + \ldots + \beta_n \; x_n^n + \epsilon$$

And we can discover the best possible values for vector β by using Eq. 5,

$$\widehat{\beta} = \left(x^T x\right)^{-1} x^T y$$

In this case, β will be more promising on large datasets rather than small datasets (the available data of COVID-19 is not enough yet) because this method failed to quantify the certainty [34]. Here, we need to make little change with β, determine a distribution instead of a single point estimation and it is all that Bayesian Ridge Regression does in this model. Now, when β is a distribution instead of a mere number our dependent variable ($\widehat{y} = f(X)$) also turns into stochastic and becomes a distribution too.

$$\beta_0 + \beta_1 x_1^1 + \beta_2 x_2^2 + \ldots + \beta_n \; x_n^n + \epsilon \rightarrow \widehat{y}$$

This means that we have confidence interval in our prediction and it became necessary to encounter uncertainty in the case of COVID-19 progression forecasting when datasets are rapidly growing but not sufficient yet. Besides, in Eq. 4 of the model, we also used L^2 (Ridge) regularization to makes model less prone to overfitting.

$$\beta = L(y_i, x_i) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \sum_{i=1}^{n} \beta_i^2$$

Ridge regression is better to use when all the weights are equal sizes and the dataset has no outliers.

Table 4 Best Fitted Parameters for Model

Country	Best Fitted Parameters				
	tol	alpha_1	alpha_2	lambda_1	lambda_2
Italy	0.0001	1.00E-06	0.0001	0.0001	1.00E-07
U.S.	0.01	0.0001	0.0001	1.00E-06	0.0001
Spain	0.0001	1.00E-07	1.00E-06	0.0001	1.00E-07

3 Significance of proposed model in COVID-19 outbreak

Clinical trials and diagnosis are very expensive and their outcomes are crucial to the concerned stakeholders and, hence, there is considerable pressure to optimize them. In medical treatments, clinicians and nurses very often have to make various complex and critical decisions during the diagnosis of the patients. In reality, these decisions are full of uncertainty and unpredictability. However, based on the available information, obtained from various clinical and diagnostic tests and situation of the patient, both clinicians and nurses try to reduce their uncertainty in clinical decisions and attempts to shift to

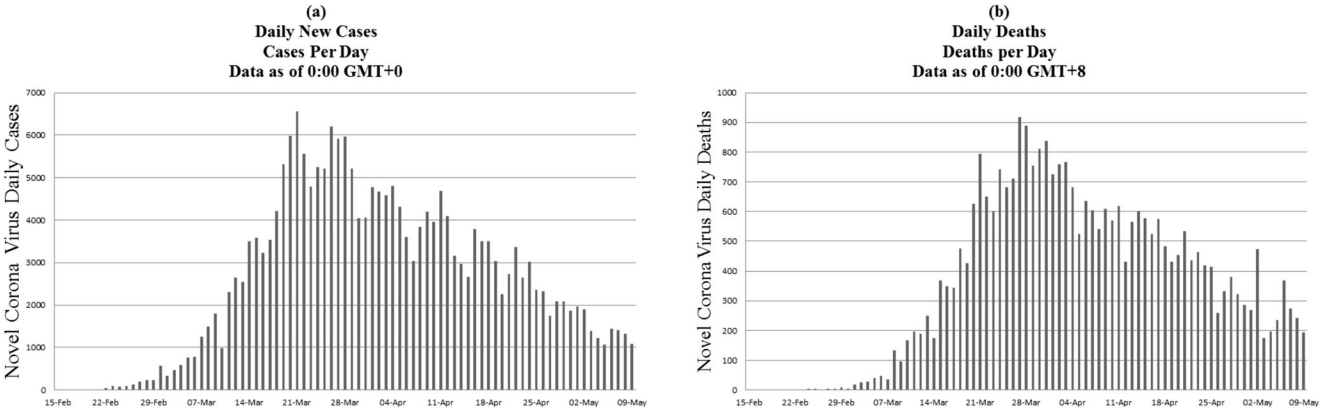

Fig. 2 Daily Death and New Cases in Italy

the predictability of the chance of improvement in patient's condition. In the case of the COVID-19 pandemic, the situation is the same as any other clinical trials. Many pre-planning and controlling need good prediction for the magnitude of infected cases as well as the measurement of uncertainty. One route of optimization is to make better use of all available information, and Bayesian statistics provides this opportunity. Bayesian statistics provide a formal mathematical method for combining prior information with current information at the design stage, during the conduct of the trial, and at the analysis stage. The main reason for using a Bayesian approach to COVID-19 is that it facilitates representing and taking fuller account of the uncertainties related to models and parameter values. In contrast, most decision analyses based on maximum likelihood (or least squares) estimation involve fixing the values of parameters that may, in actuality, have an important bearing on the outcome of the analysis and for which there is considerable uncertainty. One of the major benefits of the Bayesian approach is the ability to incorporate prior information.

Bayesian inference based approach is really important to conduct for COVID-19 pandemic rather than doing point estimations because it makes it possible to obtain probability density functions for model parameters and estimate the uncertainty that is important in the risk assessment analytics. In the Bayesian regression approach, we can take into account

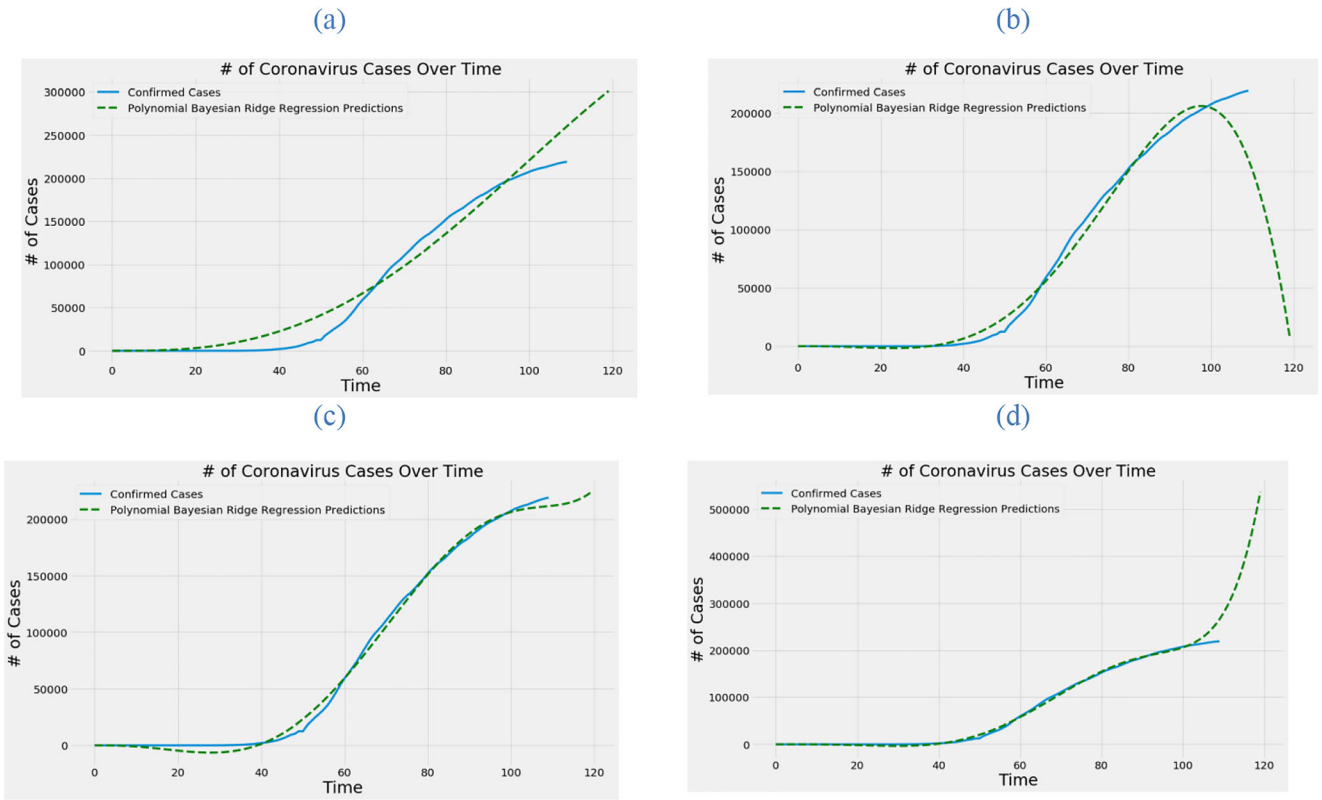

Fig. 3 Italy Cases forecasting (**a**) Represent degree-4, (**b**) degree-5, (**c**) degree-6, and (**d**) degree-7 PBRR as well

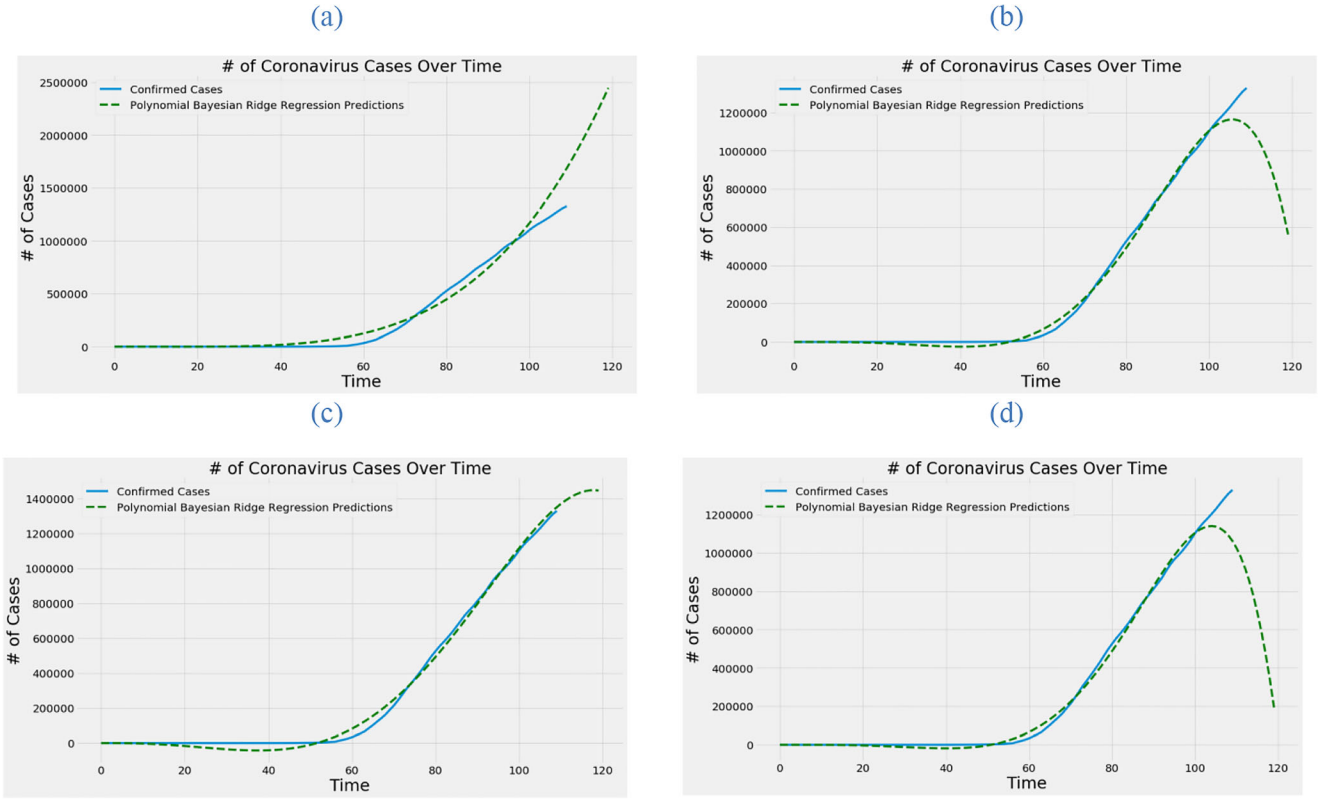

Fig. 4 U.S. Cases forecasting (**a**) Represent degree-4, (**b**) degree-5, (**c**) degree-6, and (**d**) degree-7 PBRR as well

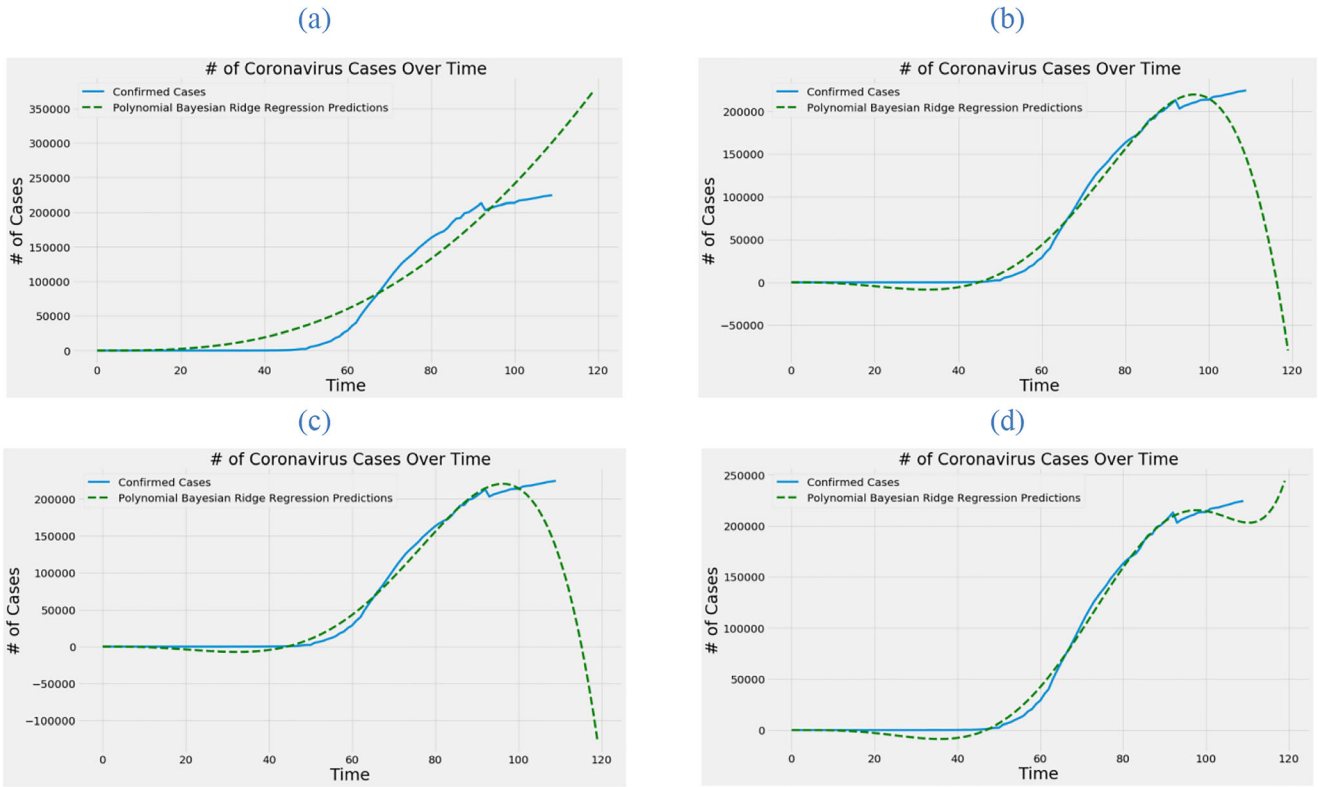

Fig. 5 Spain Cases forecasting (**a**) Represent degree-4, (**b**) degree-5, (**c**) degree-6, and (**d**) degree-7 PBRR as well

Table 5 Comparisons of Predicted and Actual Number of Cases

Date	Spain		Italy		U.S	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
06-05-2020	212,962	220,325	211,379	214,457	1,147,912	1,233,527
07-05-2020	212,500	221,447	212,079	215,858	1,165,587	1,261,409
08-05-2020	212,274	222,857	212,760	217,185	1,177,504	1,288,587
09-05-2020	212,391	223,578	213,448	218,268	1,183,026	1,314,320
10-05-2020	212,973	224,350	214,173	219,070	1,184,105	1,334,084
11-05-2020	214,154	227,436	214,969	219,814	1,185,812	1,352,962

expert opinions via information prior to distribution. Other models are developed with good accuracy but as well as data become available, those entire algorithms will not able to survive without a few evaluations due to the dynamic nature of pandemic escalation of the COVID-19 but the proposed model corrects the distributions for model parameters and forecasting results using parameters distributions. This approach has always been used for pandemic and clinical forecasting due to uncertainty measurement for example in [21] Bayesian modeling approach has used to calculate vaccine efficacy in the declining Ebola epidemic and [23, 24] demonstrated a Bayesian scheme for emerging infectious diseases and show how to use such inferences to formulate significance tests on future epidemiological observations. In short, Bayesian methods have the following advantages [35] over other time-series machine learning approaches:

- It provides an organized way of combining prior information with data, within a solid decision theoretical framework.
- It's an inference based learning approach based on previously available data without reliance on asymptotic approximation and such learning gives the consistency of the results with a small sample and large sample equally.
- It is based on the likelihood principle which gives identical inferences with distinct sampling designs.

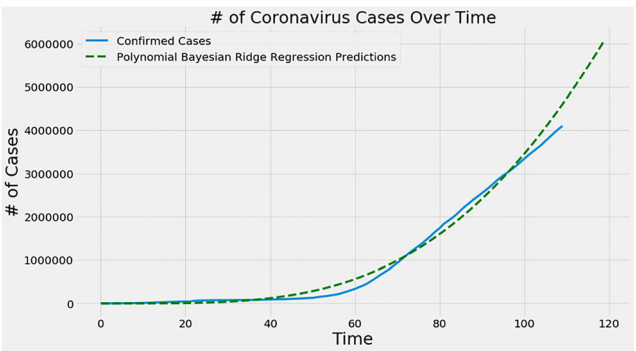

Fig. 6 Word-Wide Cases Forecasting

- Interpretability of distribution of various parameters used in the model.

4 Case studies

The method of the present study is unique because the model uses prior and posterior distribution to estimate the confirmed cases. The model should not only judge by the accuracy but also on the reliability of the prediction it makes using prior and posterior knowledge fetched from the data. To test the results and get the accuracy of the model we have proposed a case study of three countries- the U.S, Spain, and Italy. We implemented the proposed model with hybridization of polynomial fitting of degree 4, 5, 6, and 7 because we have observed the best estimation are happen within this range of degrees.

4.1 Italy cases

The confirmed cases in Italy were the lowest since 13th March but the deaths remain stubbornly high, have hovered between 600 and 800 for the last few weeks (See Fig. 2).

Using PBRR, we fitted polynomials and discovered that degree-6 is the best fit for the dataset of Italy. In Fig. 3 the solid blue line demonstrating the actual confirmed cases and dashed green line represented observation calculated by the model. In Fig. 3 (a), the degree-4 PBRR is suffering from underfitting and poorly estimate the cases for the unseen days. Also, in Fig. 3 (b), the model showing overfitting and overestimate on the testing data. Fig. 3 (b), sudden decrement in the number of cases which is not a good prediction considering the ongoing situation of Italy. Fig. 3 (c), degree-6 PBRR given RMSE 418.36 with an accuracy of 91% on testing data.

4.2 U.S. cases

In Fig. 4 we have plotted four polynomial-curves using PBRR of different degrees and observed degree-6 is well suited the

Table 6 Models Comparisons

S.R	Model	RMSE	Accuracy	Prior Distribution/ Posterior Distribution (For parameters)	SD
1	Regression [36]	1.75	–	No	–
2	Bayesian Linear Regression	0.2	82	Yes	0.012
3	MultiOutputRegressor + XGBoost [16]	–	82.4	No	–
4	SEIR [36]	1.52	–	No	–
5	ARIMA [37, 38]	0.08	93.75	No	–
6	Prophet [38]	0.06	–	No	–
7	NBEATS [38]	0.05	–	No	–
8	Gluonts [38]	0.10	–	No	–
9	LSTM [39]	–	92.67	No	–
10	Proposed Model (PBRR)	0.04	91	Yes	0.003

case examined using Root Mean Square Error (RMSE). We have recorded 723.75 RMSE with an accuracy of 88% after training and testing our model on the U.S. dataset. Within 60 days of the first COVID-19 positive case occurrence, the number of confirmed cases started growing exponentially. In Fig. 4 (a) 4-degree PBRR fitted but given poor performance-tested data due to overfitting on training data. Also, in Fig. 4 (b) 5-degree and (d) 7-degree PBRR fitted so well but after 100 days, it started decreasing which is not suitable for the current circumstances.

4.3 Spain cases

Through our investigation on the dataset of Spain, an instant decrement is recorded on the 95th days of first case arrival. Similar to the previous we fitted four different PBRR on Spain dataset too and found that 7-degree is the best fit that not correctly estimates confirmed cases for unseen days but also tracks the decrement happen earlier (See Fig. 5). The model, in Fig. 5 (a) is underfitting that neither predicts the unseen observation nor performed well on the training dataset. The other two models (Fig. 5 (b) and (c)) are not suitable for the present ongoing. The model, in Fig. 5 (d) estimates the testing data having RMSE 624.27 with an accuracy of 90.5%.

In the above case studies, we have fitted and found different parameters for the predictions (Table 4). Now, we can predict for the upcoming days. So, we have predicted for 6 days and compare with the actual number of cases on those days (Table 5).

5 Results and discussions

It is demanding to construct a model to predict the dynamic progression of COVID-19 situations. So many researchers are struggling to find and implementing such models with optimal parameters and unknown variables which lead them to uncertainty. PBRR model is different from all the studies published or at least discussed in the literature survey because of its nature of making an estimation. It is a complex mathematical model that more focused to discover distribution instead of making a single value linear prediction of the dependent variable and this feature makes it more promising.

As far as we have seen in all the above-mentioned case studies different polynomial based on Bayesian belief having a range of degrees between 4 and 7 best fit and enable us to forecast future infected cases of COVID-19. Instead of applying any specific country, we can also estimate the cases on the worldwide dataset. Fig. 6, demonstrates the curve fitting using the PBRR model of degree 5 on world-wide data with accuracy 89% on testing data. We also estimate the magnitude of confirmed cases in the upcoming 10 days. Applying PBRR on world-wide data is means scaling the independent variables but our model also survived in this scenario and showing the consistency of the system.

To prove the novelty and superiority of the proposed model, we have compared several models (Table 6) based on many attributes which are following-

- Root Mean Square Error (RMSE)
- Accuracy of Prediction
- Prior Distribution/ Posterior Distribution (For parameters)
- Standard Deviation of Prediction (SD)

After the comparison, we finally observed that the proposed model is better than other in the term of RMSE and comparable equal in term of accuracy with ARIMA and LSTM. Although, ARIMA and LSTM are giving little bit more accurate results PBRR using the prior and posterior distribution for the model parameters which is not used by any of either ARIMA or LSTM. We also experiment with Bayesian

Linear Regression with using the prior and posterior distribution for the model parameters which has not given satisfactory result in the term of RMSE, accuracy, and SD. In section 2, we have already discussed the importance of the prior and posterior distribution for the model parameters.

No doubt, LSTM is a deep learning based advanced approach to forecast time series data but it also has some drawbacks compare to proposed model e.g. longer time to train, more memory, overfitting, sensitive to different random weight initializations etc. The overfitting is one of the major issues of the LSTM which has overcome in proposed model by adding ridge regularization. We have a sequential path from older past cells to the current one in LSTM hidden layers. In fact the path is now even more complicated, because it has additive and forget branches attached to it. LSTM and GRU and derivatives are able to learn a lot of longer term information but they can remember sequences of 100 s, not 1000s or 10,000 s or more as given here [40]. Moreover, RNNs are not hardware friendly. It takes a lot of resources we do not have to train these networks fast. Also it takes many resources to run these models in the cloud, the cloud is not scalable [40].

6 Conclusion

PBRR modeling not only has sufficient accuracy but also reliable than other methods. In present circumstances when thousands of people are losing their loving ones or own lives a model with more promising algorithms is needed along with good accuracy. Prediction with misconceptions may lead to a serious problem for health care professionals as well as governments. Although, PBRR is giving reliable results the reality is the forecasting of any pandemic is not only merely dependent on previous observations or time-series analytical inference. Many more important factors influence the magnitude of infection like healthcare system stability, education, awareness of people, weather, lockdown, and social-distancing, etc. Soon, the researcher may come up with different robust models that also consider these factors.

References

1. "World Health Organization. Novel coronavirus - China.," Available from http//www.who.int/csr/zxcvXDdon/12 -january-2020-novel-coronavirus-china/en/, accessed 21 April. 2020
2. "World Health Organization," https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it
3. "Ministry of Health & Family Welfare, Government of India.," COVID-19 India Updat. https//www.mohfw.gov.in/ , accessed 21 April. 2020
4. Basing A, Tay S (2014) Malaria transmission dynamics of the anopheles mosquito in kumasi, ghana. International J. Infect Dis Ther 21:22
5. Cushing JM, Hyman JM (2006) Bifurcation analysis of a mathematical model for malaria transmission. SIAM J Appl Math 67(1): 24–45
6. Sharomi O, Podder CN, Gumel AB (2008) Mathematical analysis of the transmission dynamics of HIV/TB coinfection in the presence of treatment. Math Biosci Eng 5(1):145–174
7. Vellingiri B, Jayaramayya K, Iyer M, Narayanasamy A, Govindasamy V, Giridharan B, Ganesan S, Venugopal A, Venkatesan D, Ganesan H, Rajagopalan K, Rahman PKSM, Cho SG, Kumar NS, Subramaniam MD (2020) COVID-19: a promising cure for the global panic. Sci Total Environ 725:138277
8. Shah K, Alqudah MA, Jarad F, Abdeljawad T (2020) Semi-analytical study of Pine Wilt Disease model with convex rate under Caputo–Febrizio fractional order derivative. Chaos, Solitons & Fractals 135:109754
9. Jajarmi A, Yusuf A, Baleanu D, Inc M (2020) A new fractional HRSV model and its optimal control: a non-singular operator approach. Phys A Stat Mech its Appl 547:123860
10. Jain R, Sontisirikit S, Iamsirithaworn S, Prendinger H (2019) Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. BMC Infect Dis 19(1):272
11. W Naudé (2020). "Artificial intelligence vs COVID-19: limitations, constraints and pitfalls," AI Soc
12. Tomar A, Gupta N (2020) Science of the Total environment prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci Total Environ 728:138762
13. R Ranjan (2020). "Predictions for COVID-19 outbreak in India using Epidemiological models predictions for COVID-19 outbreak in India using," no. March
14. V Kumar, R Chimmula, and L Zhang (2020). "Time Series Forecasting of COVID-19 transmission in Canada Using LSTM Networks," Chaos, Solitons Fractals Interdiscip. J. Nonlinear Sci. Nonequilibrium Complex Phenom., p. 109864
15. R Sujatha, J Chatterjee, and A Ella Hassanien (2020). "A machine learning methodology for forecasting of the COVID-19 cases in India," TechRxiv. Prepr
16. Y Suzuki and A Suzuki (2020). "Machine learning model estimating number of COVID-19 infection cases over coming 24 days in every province of South Korea (XGBoost and MultiOutputRegressor)," medRxiv, p. 2020.05.10.20097527
17. Z Yang et al. (2020). "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," J Thorac Dis, vol. 12, no. 3
18. C Nanda et al. (2020). "Forecasting COVID-19 impact in India using pandemic waves Nonlinear Growth Models," no. April
19. MK Arti (2020). "Modeling and Predictions for COVID 19 Spread in India," no. April
20. "Bayesian linear regression," Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Bayesian_linear_regression
21. A Camacho, RM Eggo, S Funk, CH Watson, AJ Kucharski, and WJ Edmunds (2015). "Estimating the probability of demonstrating vaccine efficacy in the declining Ebola epidemic: a Bayesian modelling approach," BMJ Open, vol. 5, no. 12
22. WA Link and RJBTBI Barker, Eds., (2010). "Chapter 5 - Bayesian Prediction," London: Academic Press, pp. 77–107
23. Jewell CP, Kypraios T, Neal P, Roberts GO (2009) Bayesian analysis for emerging infectious diseases. Bayesian Anal 4(3):465–496
24. LMA B, RM R (2008) Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases. PLoS One 3(5):2185
25. Sebastiani P, Mandl KD, Szolovits P, Kohane IS, Ramoni MF (2006) A Bayesian dynamic model for influenza surveillance. Stat Med 25(11):1803–1816

26. D Foley (2018). "A Bayesian Approach to Time Series Forecasting, " [Online]. Available: https://towardsdatascience.com/a-bayesian-approach-to-time-series-forecasting-d97dd4168cb7. [Accessed: 27-Jun-2020]
27. "COVID-19 Datasets from Johns Hopkins University." [Online]. Available: https://github.com/CSSEGISandData/COVID-19. [Accessed: 11-May-2020]
28. Xu JLY, Weaver JB, Healy DM (1994) Wavelet transform domain filters: a spatially selective noise filtration technique. IEEE Trans Image Process 3(6):747–758
29. Hastie, T, Tibshirani, R, J Friedman (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer
30. "Bayesian Linear Regression Models with PyMC3," *Quantstart*. [Online]. Available: https://www.quantstart.com/articles/Bayesian-Linear-Regression-Models-with-PyMC3/
31. M Gruber (1998). "Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators.," Boca Rat CRC Press, no. ISBN 0–8247–0156-9, pp. 7–15
32. Pedregosa F et al (2011) Scikit-learn: machine learning in {P}ython. J Mach Learn Res 12:2825–2830
33. Lunn DSD, Jackson C, Best N, Thomas A (2012) The BUGS book: a practical introduction to Bayesian analysis, ser. Taylor & Francis, Chapman & Hall/CRC Texts in Statistical Science
34. MCH Dan Lu,Ming Ye (2012). "Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification," Water Resour. Res., vol. 48
35. "Introduction to Bayesian Analysis Procedures," SAS/STAT 14.3 User's Guide. [Online]. Available: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_introbayes_sect001.htm. [Accessed: 06-Jul-2020]
36. G Pandey, P Chaudhary, R Gupta, and S Pal (2020). "SEIR and regression model based COVID-19 outbreak predictions in India
37. Chintalapudi N, Battineni G, Amenta F (2020) COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: a data driven model approach. J Microbiol Immunol Infect 53(3):396–403
38. Papastefanopoulos V, Linardatos P, Kotsiantis S (2020) COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population. Appl Sci 10(11):3880
39. Chimmula VKR, Zhang L (2020) Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals 135:109864
40. E Culurciello, "The fall of RNN / LSTM." [Online]. Available: https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0. [Accessed: 20-Aug-2020]

Mohd Saqib received his B.Sc. in Computer Application from the Department of Computer Science, Aligarh Muslim University (A.M.U.), Aligarh, India in 2015. And also completed a Master of Computer Application and Science from the same university with distinction in 2018. He was a research assistant in the Centre of Advanced Research in Electrified Transportation (CARET), A.M.U and during this period, he worked on projects related to monitoring and accessing smart grid data via cloud computing. Nowadays, he is pursuing M.Tech in Data Analytics from the IIT (ISM) Dhanbad. He has been involved in automation in smart grid and applied artificial intelligence in seismology. His area of interest is data analysis, artificial intelligence, and applied statistics.