Prediction and forecasting of worldwide coronavirus (COVID-19) outbreak using time series and machine learning

Priyank Jain1 | Shriya Sahu2

1 Indian Institute of Information Technology, Bhopal, Madhya Pradesh, India
2 Atal Bihari Vajpayee University, Bilaspur, Chhattisgarh, India

Summary
How will the newly discovered coronavirus (COVID-19) affect the world and what will be its global impact? For answering this question, we will require a prediction of overall recoveries and fatalities, as well as a reliable prognosis of coronavirus cases. Predicting, however, requires an ample total of past data related to it. On any particular day, the prediction is unclear since events in the future rarely repeat themselves the way that they did in the past. Furthermore, forecasts and predictions are determined by the absolute interests, accuracy of the data, and prophesied variables. In addition, psychological factors play an enormous role in how people perceive and react to the danger from the disease and therefore the fear that it is going to affect them personally. This research paper advances an unbiased method for predicting the increase of the COVID-19 employing a simple, but powerful method to do so. Assumed that the data are accurate and reliable which the longer term will still follow an equivalent disease pattern, our projections intimate with a large association. Within the COVID-19 cases were documented, in contingency, there is a steady increase. The hazards are far away from symmetric, as underestimating a pandemic’s spread and failing to do enough to prevent it is far a lot worse than overspending and being too cautious when it will not be needed. This paper illustrates the timeline of a live forecasting study with huge implied implications for devising and decision-making and gives unbiased predictions on COVID-19 confirmed cases, recovered cases, deaths, and ongoing cases are shown on a continental map using data science and machine learning (ML) approaches. Utilizing these ML-based techniques, the proposed system predicts the accurate COVID-19 cases and gives better performance.

KEYWORDS
coronavirus (COVID-19), curve fitting, data visualization, infodemic, machine learning classifier, SARS-CoV-2, time series

1 INTRODUCTION

COVID-19 is defined by the WHO, and it is a new Corona Virus Disease that became an epidemic on 3/11/2020 (mm/dd/yy). The coronavirus is not new. Several coronaviruses have caused epidemics in the past, most notably SARS and MERS. The WHO announced an outbreak of “unknown cause pneumonia” in Wuhan City, China, on 12/31/2019. Therefore, this COVID 19 badly affects many other countries like Iran, Spain, the United States, and Italy. The coronavirus made its first appearance in India on January 30, 2020. COVID-19 transmission is known by contact, fomites, and respiratory droplets. In order to take the required precautions in terms of mitigation efforts and preparedness strategies, forecasting the Covid-19
epidemic size on a global and regional scale is critical. In order to take required actions, it is especially important to forecast the number of cumulative and daily cases, and deaths. In the contest against the pandemic, the capacity to determine the rate at which the disease is spreading is critical. Furthermore, to aid governments in policy development and public health planning to handle the pandemic’s consequences, being aware of the extent of the pandemic’s spread at any particular time is important. Apart from this, Falco oral transmission and food transmission are considered routes of transmission. Due to these modes of transmission, the COVID-19 disease is spreading rapidly.

In the beginning, it is symptomatic and its prominent symptoms are cough, very high fever, and difficulty in respiration. After the incubation phase, the Covid-19 disease symptoms developed to appear which lasts about 5 days. Covid-19 has starting with the first disease symptoms and some end with the patient’s death, in the middle of 14 days Covid-19 has a cycle of 6–41 days. Due to this phase, the immune system state and the patient’s age are particularly important parameters. Now it becomes an asymptomatic disease also and in today’s scenario, asymptomatic cases are much more than symptomatic ones. The recommending ways to deal with it and the disease’s diagnosing capabilities are improved by the three-pronged strategy, which is launched by WHO. The virus is deadly, fast, and furious and is swiftly spreading over the world. In order to contain the virus and prevent it from spreading, the Indian government enforces rules and restrictions. Each individual has taken preventative measures, resource allocation, and healthcare needs are evaluated by assessing and predicting how the virus is spreading throughout the community is important. All of these variables combine to make COVID-19 prevention and treatment difficult. The condition primarily affects the respiratory systems. The lungs are the primary organs that are impacted by this condition. According to the SARS (severe acute respiratory syndrome) incubation durations and MERS (Middle East respiratory syndrome), the infected person develops symptoms within 2–14 days. As a result, patients are at significant risk of death. Based on the identified symptoms, medications are given to patients for this condition.

COVID-19 prevention methods include maintaining social distance, washing hands frequently, preventing contact with infectious people, wearing a mask, and avoiding touching the mouth, nose, and face. The government advise the individuals to stay at home and suspended all forms of public transportation. To prevent COVID-19 transfer from one person to another, maintaining social distance is the most effective strategy due to the lack of a viable cure for this disease. The lockdown curfews and quarantines are China’s model that can be adopted by other countries. They also restricted civilian movement, except distribution of important goods and services, and medical emergencies. In the places with higher temperatures and humidity, the virus will spread more slowly than in areas with average records as hypothesized by Malik et al. The humidity and temperature thresholds have been set at 75% humidity and 15°C, respectively, based on this research. From one country to another country, the humidity and temperature factors’ thresholds differed. These numbers will be neglected from the day’s overall humidity and maximum temperature measurements. Based on accelerated genetic mutations to modify its nature and evolve new variants, more information on COVID-19 becomes accessible, and its nature and features are being found. India documented COVID-19 instances. Health officials and administration are under a significant deal of pressure to accommodate COVID-19 patients. Subsequently, to make administrative level preparations to learn about the future’s overall positive cases, employ some prediction tools.

In different fields, Machine Learning (ML)-based models have been used successfully. A reliable and accurate pandemic prediction is provided by ML-based models. Correctly allocating resources, decreasing pressure and making the situation more manageable, by assisting hospitals and healthcare management with the prediction model. Therefore, machine learning techniques and strong time series models were employed in this research to develop an analytical forecast of total positive cases for COVID-19. Also look at the progression of confirmed cases, active cases, fatalities, recoveries, and the mortality and recovery rates of COVID-19 individuals around the world by using the data of John Hopkins University. The proposed model 99% accurately predicts the new confirmed COVID-19 cases and this will facilitate the administration to make preparations accordingly to accommodate the patients. The remainder of this research is devoted as follows: In Section 2, a literature review of COVID-19 is discussed. Section 3 demonstrates the software and tools utilized in this study. Section 4 explains about research objectives. Section 5 represents the proposed methodology. Section 6 depicted the results and discussion of the study. Finally, Section 7 brings this paper to an concludes.

### 2 LITERATURE SURVEY

The literature survey is based on the analysis of prediction and forecasting of coronavirus disease throughout the world utilizing ML and other techniques. This research work utilizes ML and the time series method for prediction.

Alizadehsani et al. proposed a Semi-Supervised Classification using Limited Labeled Data (SCLLD) for automatic COVID-19 identification, which relies on Generative Adversarial Networks (GAN). They used Sobel edge detection to improve the detection accuracy of the suggested approach. Other state-of-the-art supervised approaches, like Gaussian processes, are compared to the proposed method. This is the first time a COVID-19 semi-supervised detection approach has been demonstrated to the best of our knowledge. Due to a lack of sufficient labeled data, supervised learners fail and this method can learn from a mixture of limited labeled and unlabeled data. Dash et al. For six of the world’s worst-affected countries, including India, and six of India’s high-incidence states, the Facebook Prophet method forecasts 90-day future values, containing the peak date of confirmed COVID-19 cases. The impact of the government’s measures on the infection’s rate of spread is indicated by the five important transition points in the growth curve of verified Indian cases. The chest X-ray images are used to detect COVID-19 by three uncertainty quantification strategies that are comparatively evaluated and comprehensively applied by Asgharnezhad et al. For the first time, new performance criteria
for the objective evaluation of uncertainty estimations as well as a novel concept of uncertainty confusion matrix are provided. They quantitatively demonstrate when they could trust DNN predictions for COVID-19 detection from chest X-rays using these new uncertainty performance criteria. In all classification tasks, it is worth noting that the proposed innovative uncertainty evaluation metrics are generic, and they might be used to assess probabilistic forecasts.

A COVID-19 spread is effectively predicted by the model that is presented by Sujath et al.\textsuperscript{19} Due to the disease predictions’ epidemiological example and the rate of COVID-19 instances in India, on the COVID-19 Kaggle data, they utilize vector autoregression, linear regression, and multilayer perceptron model. Based on data from Kaggle, predicted the COVID-19 cases in India. It is possible to forecast and estimate the not-too-distant future with common data on confirmed, death, and recovered cases across India over a long period. Data combination and case definition must be maintained over time for future perspective or additional assessment. Hence, Joloudari et al.\textsuperscript{20} proposed a Deep Neural Network model named DNN-GFE for COVID-19 diagnosis using CT scan images of 1229 healthy and 1252 sick data. For distinguishing between sick and healthy people, the DNN-GFE model was created to improve the accuracy of diagnostics. In order to improve COVID-19 diagnosis and a high-quality image is created using image normalization. The data is also partitioned into training, testing, and validation using the 10-fold cross-validation technique. The DNN-GFE model’s experimental results were compared to three classification models for COVID-19 diagnosis in terms of accuracy, and the DNN-GFE model had the highest accuracy.

The precision of conventional forecasting mainly depends on the data’s availability to base its predictions and evaluations of ambiguity. In outbreaks of pandemics there is no data at all at the start and then restricted as time passes, making predictions generally doubtful. On February 18, 2020, a New York Times piece warned against developing optimism about the situation, noting that the virus had been known for 60 days or more. Due to the afflicted and mortality cases being misquoted to hide the pandemic’s breadth, pandemic data is thought to be unreliable, as it was with bird flu and SARS. In the COVID-19 cases, precise estimations were not represented in the reporting, and a new category of “clinically diagnosed” was added to “lab-confirmed” cases on February 13th. Such problems reduce forecasting efficiency and enhance uncertainty, making the study of certain conclusions more complex. Associated with predicting accuracy and uncertainty, there is a major severe difficulty that has to do with the understanding of pandemics and epidemics. Lawmakers are concerned with thoughts on the steps to be taken while the common people worry about the pandemic’s impact on their lives. Moreover, pharmaceutical organizations are working on vaccines for this new virus with significant commercial importance.\textsuperscript{21} It was the situation with SARS when authorities were persuaded of the severity of the virus and got large numbers of vaccines that were never practised as its spread discontinued vaccination is not required for people. Additionally, the visualization of Covid-19 around the world is demonstrated in Figure 1.

Ayobii et al.\textsuperscript{22} forecast new cases and deaths for the next 100 days 1, 3, and 7 days ahead. This research is unique in that it conducts a full evaluation of the three deep learning algorithms listed as LSTM, convolutional LSTM, and GRU, as well as their bidirectional extensions, to forecast

![Figure 1](image-url)
COVID-19 new cases and death rate time series. To forecast new cases and fatalities in the COVID-19 time series, Bi-GRU and Bi-Conv-LSTM models have been employed. Finally, the superiority of bidirectional methods is determined and numerous error evaluation metrics are presented to compare all models. Khozeimeh et al. suggested the CNN-AE, a novel approach for predicting COVID-19 patient survival using a CNN trained with clinical data. They evaluated the method using a publicly available clinical dataset that they collected. In order to extract relevant features and compute feature correlations, the parameters of the dataset were thoroughly examined. To balance the dataset, a data augmentation approach based on autoencoders (AEs) was developed. They also tested their strategy on a different dataset to ensure its generality. The CNN-AE was compared to different pre-trained deep models that were tweaked based on CT scans to indicate that clinical data may be used for COVID-19 survival chance prediction.

Omran et al. conducted a comparison study of two deep learning techniques for forecasting COVID-19 confirmed cases and death cases. From 1/5/2020 to 6/12/2020, time-series data was processed using gated recurrent unit (GRU) and LSTM in three countries like Kuwait, Egypt, and Saudi Arabia. In Egypt and Kuwait, in death cases, the best performance is achieved by GRU and in the three countries, for the confirmed corona cases, the best performance is achieved by LSTM is depicted in this study. Sharifrazi et al. proposed a merger of a convolutional neural network (CNN), support vector machine (SVM), and Sobel filter to detect COVID-19 using X-ray pictures. To obtain the edges of the images, a new X-ray image dataset was obtained and treated to a high pass filter using a Sobel filter. The images are then loaded into a CNN deep learning model, which is then followed by an SVM classifier with a 10-fold cross-validation technique. This approach is created in such a way that it can learn with a small amount of data. The proposed CNN-SVM with Sobel filtering (CNN-SVM + Sobel) had the maximum classification accuracy of 99.02% in accurately detecting COVID-19, according to their findings. It was discovered that using the Sobel filter can increase CNN performance. Using six public databases, they also tested their developed model and found it to be the most effective.

3 | SOFTWARE AND TOOLS DESCRIPTION

3.1 | Jupyter Notebook

This Notebook is the most effective tool for producing and sharing documents, a web-based application that uses live code, equations, visualizations, and narrative text to help you with your investigation. Learn the abilities you will need to succeed in data science. Uses of the Jupyter Notebook cover data cleansing and data transformation, statistical modeling, data visualization, numerical simulation, ML and much more.

The visualization of world Covid-19 cases and the pie chart for the visualization of country-wise Covid-19 cases are demonstrated in Figures 2 and 3.

3.2 | Python IDLE

IDLE is a consolidated development environment for editing and executing python 2 versions or python 3 version programs. We obtain the output of the program as a result.

3.3 | Python

Object-oriented programming is an interpreted, open-source, and high-level language. Due to an extensive range of data science projects and applications, it is the most extensively used language among data scientists.

3.4 | Matplotlib

Static, interactive visualizations and animated were created in python from the comprehensive library of Matplotlib. The following Figure 4 illustrates the comparison graph between top countries.

3.5 | Scikit-learn

It is a Python machine learning package that is free to use. It includes scientific and numerical Python libraries like SciPy and NumPy, as well as random forests, SVM, and k-neighbor’s algorithms.
4 | RESEARCH OBJECTIVES

This step is very important because this is where the project’s purpose is planned. In this step, we predefined our objectives, that is, to aid with the analysis, prediction, and visualization of COVID-19 confirmed cases for a better deal with the global problem.

5 | PROPOSED METHODOLOGY

The current global outbreak of the novel COVID-19 has presented unexpected challenges. The global economy is frozen because Covid-19 is an extremely contagious virus. With its ability to transmit surface-to-human and human-to-human, the world is turned into a catastrophic phase. With its rapid transnational expansion fueled by increased trade and global travel, the Coronavirus, which has been connected to severe outbreaks, exacerbates public global health challenges. With the development of public health programs and disease control policies, suitability mapping of Coronavirus transmission risk is required which is especially significant in areas with medical facility shortages. Around the world, millions of people are killed by a coronavirus and this has been arresting people regularly. To prevent this communicable disease by isolating hygiene, covering your face, staying away from the community, and washing hands, but is not enough. Due to the COVID-19 cases, there are neither immunizations nor explicit antiviral medicines as per the WHO.

The spread of these illnesses can be stopped, according to the Disease Control and Prevention Centers. There are simple, easy-to-use, and cost-effective techniques that can assist reduce pollution, ranging from fundamental hand-washing principles to a group approach related to a group-based comprehensive safety strategy. Then the flow diagram of the proposed work is depicted in Figure 5. The methods used to control and prevent the infectious Covid-19 disease include hand hygiene, environmental cleanliness, patient screening and grouping, vaccination, surveillance, antibiotic management, evidence-based care coordination, appreciation of all departments that contribute to the infection avoidance plan, and a safety plan based on a comprehensive unit. There is a lot of evidence in the healthcare business that ML algorithms can create useful models to identify patients and solve problems. Grouped the individuals who are most at risk. To solve this situation, many scientists and researchers related ML techniques. Many data scientists may make the proper decisions and take precise activities to comprehend the patterns and characteristics of virus attacks. As a result, the machine learning approach and the time series model are used in this study to forecast and predict the coronavirus, respectively.

Figure 2  Visualization of world Covid-19 cases
To forecast COVID-19 positive instances, we use the raw data provided by Johns Hopkins University. Here, data is available in country-wise confirmed cases in the time-series format as explained in Section 5.2, then, we constructed a new dataset from a raw dataset which consists of two columns named dated and confirmed cases which contain dates from 01/22/2020 and cumulative cases worldwide on that date respectively as portrayed in Section 5.2. The prepared data then was used in different models. To make predictions, exponential smoothing family models were utilized. The family has given satisfying forecast accuracy over many forecasting methods and is particularly fit for small-time series. A large range of trend and forecasting patterns (like multiplicative or additive) were produced by exponential smoothing algorithms and mixtures of those based on this we chose SVR, Polynomial Regression and Bayesian Ridge which is discussed briefly in Section 5.2. We also compared and assessed the different model performance parameters for the models we have used are discussed in Section 6.1. Then we predicted confirmed cases against the test data (i.e., 30% of total data) and evaluated the optimal hyperparameters and then used the 100% data to train models with optimal hyperparameters further it is explained in section.

We concentrate on the combined daily numbers gathered internationally of the principal variable of concern: confirmed cases. Johns Hopkins University provided the information for this article. For COVID-19 case prediction, a basic time series forecasting method was used, which was proven. The data gathered from the sources contains a variety of properties, the confirmed data frame consists of a country column concerning its Covid-19 cases are tagged over time. Using this data, we projected the global Covid-19 cases. We have used three different models under machine learning and data science and for that, we have used the methodology, which is widely used by the followed data science community.

5.2 Steps of prediction

Data understanding: Data understanding depends on the previous step, that is, business understanding. The data was accumulated at this step of the method. The conclusion of what the business requires and needs will decide what data is gathered, from which sources, and by what methods. The COVID-19 Global Cases data sets were contributed by Johns Hopkins University’s (JHU) Center for Systems Science and Engineering (CSSE). Starting on January 22, 2020, the daily confirmed cases are stored in a comma-separated values (CSV) file. By taking the cumulative sum date wise of every country, a novel data frame was created with two calculated columns: date and confirmed cases, using the confirmed corona cases file to get total world cases on a particular date.
**Data preparation:** Unless it is determined that larger data is required, the data must be turned into useful subsets after it is acquired. Column verified cases will be utilized as the dependent variable \( y \) in the above-generated data frame, with dates beginning on January 22nd will be used as explanatory variable or independent variable \( X \). \( X \) and \( y \) were reshaped into a single column NumPy array with rows equal to 185 it is a day starting from 22nd January 2020 to 24th June 2020. The model selection library was used to import model selection from which the train test split function was used to divide the present data into two categories: Training and Testing. Only 30% of the complete dataset was used for testing, while the remaining 70% was used for training. Training and Testing data scaling of independent variables was done using fit transform before using them to fit the models with these data.

**Modeling:** Once data was prepared for use, through hopefully new knowledge, whatever appropriate models, and give meaningful insights, the data must be expressed. The use of models shows patterns and compositions inside the data that give insight into the points of interest. Models were selected on a portion of the data and changes were made when needed. We will use three different models for prediction, which are as follows.

**Support vector regression:** The classification and regression procedures are performed by an SVM, which is a supervised ML approach. In a specific way (prediction, import library, fitting model, and object creation) we use it and the SVM is included in the sci-kit-learn toolkit. Then utilizes the SVM.
model to recognize patterns, signal processing, and non-linear regression because of its tendency for dealing with complex time-series, nonlinear, and dynamic data. When we use SVM for regression then these types of methods are recognized as SVR (support vector regression), it is based on the same principle as SVM in a hyperplane classification and an \( n \)-dimensional feature space is used, where features are denoted as \( n \). SVR provides us with the flexibility to determine how much error is admissible in our model and will locate a suitable hyperplane to fit the data. In most linear regression models, the squared errors are reduced as much as possible which is the goal but in contrast, SVR’s goal is to reduce the coefficients to a more precise level (Equation 1), for the L2-norm, we can state that the coefficient vector is minimized. The error expression is instead managed in the constraints, where the absolute error is set to be less than or equal to a specific margin (Equation 2), referred to as the maximum error, \( \varepsilon \) (epsilon). We can adjust epsilon to attain the desired efficiency of our model. Figure 6 shows the internal working of SVR and how the hyperplane is calculated. Our goal function and constraints are listed below:

\[
\text{Minimize} \quad \min \|w\|^2 \tag{1}
\]

\[\begin{align*}
\text{subject to} & \quad y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - \mathbf{b} \leq \varepsilon + \xi_i, \\
& \quad \mathbf{b} + \langle \mathbf{w}, \mathbf{x}_i \rangle - y_i \leq \varepsilon + \xi_i, \\
& \quad \xi_i \geq 0
\end{align*}\]
Constrains

$$|y_i - w_i x_i| \leq s$$  \hspace{1cm} (2)

where, $y_i$ is the target variable in our case, the weights or coefficients of features are described as $w_i$, the predictor (feature) is represented as $x_i$, and we have given dates as input for the prediction and it has been verified that Covid-19 cases.

We have used RandomizeSearchCv () from the model selection python library to get the optimal hyperparameters for SVR, whose values are given below.

SVR (shrinking = True, kernel = 'poly', gamma = 0.00842, epsilon = 1, degree = 4, C = 0.001)

The SVR advantage is that it provides a Kernel trick that can be utilized for representing data in a multiple plane system to get a better hyperplane here C employs a polynomial kernel with a degree equal to 4, on the training data control the misclassification cost. Figure 7 depicts the implantation of SVR on our worldwide-confirmed system. Covid-19 Cases are used to forecast confirmed cases in future.

5.2.1 Polynomial regression

The $n$th degree polynomial is modeled by the independent variable ($x$) and the dependent variable ($y$) correlation and it is a linear regression. Between the $x$ value and the $y$ conditional mean, a nonlinear relation is fitted by polynomial regression, indicated by $E(y - x)$. Here, $y$ signifies confirmed Covid-19 in our situation and $x$ signifies the dates, while cases specify the cases. The polynomial equation of degree $n$ is used in polynomial regression, which is written as:

$$y = d_0 + d_1 x + d_2 x^2 + d_3 x^3 + \cdots + d_n x^n$$  \hspace{1cm} (3)

**Figure 7** Illustrative example of simple SVR
where, \(d_0\) is the bias, \(d_1, d_2, d_3, \ldots, d_n\) are the weights of the polynomial regression equation, and \(n\) is the degree of the polynomial regression. A simple example of how polynomial regression represent data is portrayed in Figure 8. In our model, we have used \(n\) equals 2. The above equation can be stated as follows for our model:

\[
y = d_0 + d_1x + d_2x^2
\]  

(4)

The weights or coefficients of the above equation were computed by fitting the data as 309145.03705541, \(-33219.73886033, 631.54498233\) for \(d_0, d_1,\) and \(d_2\), respectively. The above equation after putting these values is:

\[
y = (309145.03705541) + (-33219.73886033)x + (631.54498233)x^2
\]  

(5)

5.2.2 Bayesian ridge regression

Bayesian regression al-to survive poorly distributed data, it lows a natural process that employs probability distributions rather than point estimations to construct linear regression. Rather than being approximated as a single value, the output or response “\(y\)” is considered to be chosen from a probability distribution. We have used it as it provides fewer errors in cases where prediction is uncertain. The Covid-19 confirms cases prediction is very uncertain so using Bayesian with a ridge will reduce the uncertainty for prediction. Mathematically, the response \(y\) is presumed to get a fully probabilistic mode that to be Gaussian distributed around \(X_w\) as follows

\[
p(\,y \mid X, w, a) = N(y | X, a)
\]  

(6)

So here Bayesian Ridge regression is utilized, which calculates a probabilistic model of the regression issue, it is the common practical type of Bayesian regression. Spherical Gaussian gives the prior coefficient \(w\)’s.

\[
p(\,w \mid \lambda) = N(w | 0, \lambda^{-1}I_p)
\]  

(7)

The above-resulted model is known as Bayesian Ridge Regression. A flowchart depicting the model’s operation is represented in Figure 9.

We have used RandomizeSearchCv() from model selection library to get the suitable hyperparameters for Bayesian Ridge, whose values are given below:

BayesianRidge(alpha 1 = 1e−06, alpha 2 = 1e−06, compute score = False, copy X = True, fit intercept = False, lambda 1 = 1e−06, lambda 2 = 1e−06, n iter = 300, tol = 0.00, 1normalize = False, verbose = False).

alpha 1 is the hyperparameter for the shape parameter’s Gamma distribution, it takes precedence over the alpha value.

alpha 2 is the Gamma distribution’s inverse scaling parameter before the alpha parameter and it is a hyperparameter.

Lambda 1 is a Gamma distribution shape parameter before the lambda parameter and it is a hyperparameter.

Lambda 2 is a hyperparameter related to the lambda parameter and it is a Gamma distribution inverse scale parameter.

\[
y = b_0 + b_1x_1 + b_2x_1^2
\]  

FIGURE 8 Flow chart for modeling and prediction using SVR
RESULTS AND DISCUSSION

In this study, the prediction and forecasting of Covid are evaluated using three methods like support vector machine, polynomial regression, and Bayesian ridge regression models. These three models predict the coronavirus cases. The data for Covid cases are from 25th July 2020 to 8 August 2020. In this study, the predictions occurred for these data sets, using Jupyter Notebook, Python IDLE, Python, and Matplotlib software, respectively.

6.1 Evaluation

The chosen model must be tested. It is normally done by having a pre-selected test, established to run on the trained model. It will allow us to assess the model’s performance on a set that it considers to be unique. Results from this are adopted to define the effectiveness of the model and foreshadow its part in the next and last stage. We calculated the optimal model parameters in the above sections, we will use them in this section and at this, the trained model will be evaluated against 30% of the overall predefined test data. Predicted values and graphical representation of test data are shown in Figures 9–11. Against the test data, testing the model after that 100% data was utilized to train the models.

The graphical representation of test data and predicted values for support vector regression is depicted in Figure 10. The model performance parameters are evaluated and compared to other ways. Table 1 displays, as their Mean Absolute Error (MAE), Mean Squared Error (MSE), and $R^2$ values, as well as a variety of models, were displayed in Table 1.

6.2 Prediction

This stage entails applying the model to work on data that is not included in the dataset. The new intercommunications at this phase might explain the new variables and requirements for the dataset and model. These new difficulties could start a revision of both problem statement and operations, or the model and data, or both. The below-mentioned Figure 11 represents the test data and predicted values for Bayesian ridge regression. Here...
we have got the predicted time-series using different models separately for 10 more days that start just after 23/07/2020 or 186th day from the starting, that is, 22/01/2020. The graphical representation for test data and predicted values for polynomial regression is demonstrated in Figure 12. Tables 2–4 represent the predicted total persons affected and confirmed by COVID-19 cases using SVR, Polynomial Regression and Bayesian Ridge Regression, respectively. The plots of previous and predicted values for SVR are predicted in Figure 13. Then 10 more days extended this dotted line, which represents the prediction of those days. Figure 14 depicts a plot of past and forecasted values for Polynomial Regression, with the dotted line extended for an additional 10 days which represents the prediction of those days and Figure 15 shows a plot of past and predicting values for Bayesian Ridge Regression, with the dotted line extended to 10 more days, representing the prediction for those days.

Adaptability is required at each step along with communication to keep the project on track. During the procedure, it may be essential to return to a previous step and make modifications at any point. The key point of this method is that it is cyclical; hence, even at the finish, you are having extra business understanding encounter to consider the viability after deployment. The journey stretches.

By analyzing and fitting the data Covid-19 data into different models and after visualizing their results were pretty guessable. We have included the results obtained from experiments utilizing several ML techniques in detail and simultaneously, we have also examined the

TABLE 1 Evaluation of different models by comparing different statistical measures

| Model                      | (MAE) × 10⁻⁶ | (MSE) × 10⁻⁶ | R²   |
|----------------------------|--------------|--------------|------|
| Support vector machine (SVM) | 1.1877       | 1758168.256  | 0.927924 |
| Polynomial regression      | 0.48059      | 280191.562   | 0.991420  |
| Bayesian ridge regression  | 0.45813      | 256982.568   | 0.992056  |

FIGURE 10 Graphical representation of test data and predicted values for support vector regressions

FIGURE 11 Graphical representation of test data and predicted values for Bayesian ridge regression
FIGURE 12  Graphical representation of test data and predicted values for polynomial regression

TABLE 2  SVM's predicted versus actual cases

| Date       | SVM predicted | Actual cases | Error % |
|------------|---------------|--------------|---------|
| 07/25/2020 | 18183644.0    | 16,194,795   | 12.28%  |
| 07/26/2020 | 18578544.0    | 16,420,090   | 13.14%  |
| 07/27/2020 | 18979866.0    | 16,386,648   | 15.88%  |
| 07/28/2020 | 19387678.0    | 16,887,096   | 14.80%  |
| 07/29/2020 | 19802049.0    | 17,176,151   | 15.28%  |
| 07/30/2020 | 20223051.0    | 17,463,557   | 15.80%  |
| 07/31/2020 | 20650753.0    | 17,753,101   | 16.32%  |
| 08/01/2020 | 21085225.0    | 18,011,723   | 17.28%  |
| 08/02/2020 | 21526539.0    | 18,232,101   | 18.28%  |
| 08/03/2020 | 21974767.0    | 18,483,206   | 17.28%  |

TABLE 3  Polynomial regression predicted versus actual cases

| Date       | Polynomial predicted | Actual cases | Error % |
|------------|----------------------|--------------|---------|
| 07/25/2020 | 15778120.0           | 16,194,795   | 2.57%   |
| 07/26/2020 | 15979204.0           | 16,420,090   | 2.68%   |
| 07/27/2020 | 16181550.0           | 16,386,648   | 1.25%   |
| 07/28/2020 | 16385160.0           | 16,887,096   | 2.97%   |
| 07/29/2020 | 16590033.0           | 17,176,151   | 3.41%   |
| 07/30/2020 | 16796169.0           | 17,463,557   | 3.80%   |
| 07/31/2020 | 17003567.0           | 17,753,101   | 4.22%   |
| 08/01/2020 | 17212229.0           | 18,011,723   | 4.42%   |
| 08/02/2020 | 17422154.0           | 18,232,101   | 4.44%   |
| 08/03/2020 | 17633343.0           | 18,483,206   | 4.5%    |

Performance results of our proposed models Support Vector Regression, Polynomial Regression, and Bayesian Ridge Regression. The overall recoveries, positive cases, and fatalities from 01/22/2020 (Day 1) to 07/24/2020 (Day 185) were included in the dataset used to estimate COVID-19 confirmed cases in different countries/regions, Country/Region names, and dates. The predicting confirmed cases dates and total positive cases knowledge were used.
TABLE 4 Bayesian ridge predicted versus actual cases

| Date      | Bayesian predicted | Actual cases | Error % |
|-----------|--------------------|--------------|---------|
| 07/25/2020| 15720201.0         | 16,194,795   | 2.93%   |
| 07/26/2020| 15920251.0         | 16,420,090   | 3.04%   |
| 07/27/2020| 16121556.0         | 16,386,648   | 1.61%   |
| 07/28/2020| 16324114.0         | 16,887,096   | 3.33%   |
| 07/29/2020| 16527928.0         | 17,176,151   | 3.78%   |
| 07/30/2020| 16732995.0         | 17,463,557   | 4.18%   |
| 07/31/2020| 16939318.0         | 17,753,101   | 4.58%   |
| 08/01/2020| 17146894.0         | 18,011,723   | 4.80%   |
| 08/02/2020| 17355725.0         | 18,232,101   | 4.80%   |
| 08/03/2020| 17565811.0         | 18,483,206   | 4.90%   |

FIGURE 13 SVM prediction

FIGURE 14 Polynomial regression predictions

The global increase rate of confirmed cases is demonstrated in Figure 1. Official data gathering began on January 22, 2020, and most of the world went into lockdown in the first week of March 2020, which does not appear to decrease the total confirmed cases in Figure 1. However, at the top, the curve’s growth rate of verified cases has begun to increase more exponentially, which is a bad sign.

- The predictions given by the SVM’s SVR (shown in Figure 13) with hyper-parameter values of gamma equal to 0.00842, epsilon is equal to 1, C was given 0.001 and degree was equal to 4 all these with "Poly" kernel. The suggested SVR approach projected values for the overall cases until day 185, it gives us 92% accuracy, and on the day 185th, that is, 07/25/2020 SVR predicted cases give just 12.28% of error for actual cases. The model was predicted with an $R^2$ value equal to 0.9279, as a result, the SVR model was able to achieve excellent accuracy.
The Polynomial Regression predictions shown in Figure 14 with degree 2 have good efficiency. Until day 185, the detected values for the entire cases with the proposed Polynomial Regression model gave us 99% accuracy, and on day 185th, that is, 07/25/2020 SVR predicted cases give just 2.57% of error concerning actual cases. The $R^2$ is 0.9914. After Bayesian Ridge Regression, this gave us the second-best accuracy.

The Bayesian Ridge Regression’s forecast is shown in Figure 15, it has the maximum accuracy rate. The hyper-parameters used for it are "alpha 1": 0.001, "alpha 2": 1e−09, "lambda 1": 1e−08, "lambda 2": 1e−06, "normalize": False, "tol": 1e−07. Until day 185, the predicted numbers for the total cases with the proposed Bayesian Ridge Regression model gave us 99.20% accuracy, and on day 185th, that is, 07/25/2020 SVR predicted cases give just 2.93% of error for actual cases and the $R^2$ value is 0.9920.

CONCLUSION AND FUTURE SCOPE

The unknown nature of a novel coronavirus has sparked global concern, prompting a Harvard professor to predict that 40%–70% of the world’s population would be affected in the following year, according to the estimates which align with German Chancellor Angela Merkel’s warning the novel Covid-19 impact. Norman, Bar-Yam and Taleb explain the systemic risk of pandemics, the presence of fat-tailed methods due to global interconnectivity and the negatively biased approximations of mortality and spread rates. On the other hand, some argue that people are too concerned and neglect the potential, citing the new virus as the first “infodemic” as a result of today’s social media’s hyper-connectivity. The polarization of the opinions globally can be summarized by the quotes of three renowned personalities:

- "The coronavirus panic is dumb." "The coronavirus panic is dumb," said Nassim Nicholas Taleb. "I’m hoping it’s not that bad, but we should assume it until we know," says Bill Gates.

- Despite what one’s opinions are, we think that predictions and particularly in high-risk situations, its related ambiguity cloud be a significant element of the decision-making process. Apart from the important public health concerns, the risks imposed on global supply chains and the economy as a combination are also essential. People who are apprehensive about taking risks can concentrate on the worst-case possibilities and react accordingly. Choosing to reject any formal, acting conservatively, and analytical forecasts, still suggest an underlying forecasting method, even if this method (personal judgment/belief) is not formalized.

- In order to aid in the prediction of Covid-19 instances based on the data gathered, this research was performed on Covid-19. Our findings are 99% accurate, and Covid-19 is a global epidemic for which data scientists and medical professionals are still looking for a cure. Till now we have not found Abby’s cure or any 100% accurate predictor machine as this disease can be controlled by social distancing and hygiene around you which reduces the chance of getting infected.

In order to forecast the situations as quickly as feasible, we deployed three distinct Machine Learning methods. We will try to improve this project to its best in future with more resources. More neural networks, machine learning algorithms, feature engineering approaches, and deep learning will be used in the future to improve the prediction efficiency of the machine.

CONFLICT OF INTEREST
The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT
No data or models were used or generated in this study.
REFERENCES

1. Singh V, Poonia RC, Kumar S, et al. Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine. J Discrete Math Sci Cryptogr. 2020;23(8):1583-1597. doi:10.1080/09720529.2020.1784535
2. Direkglug C, Sah M. Worldwide and regional forecasting of coronavirus (Covid-19) spread using a deep learning model. bioRxiv; 2020. doi:10.1101/2020.05.23.211039
3. Papastefanopoulos V, Linardatos P, Kotsiantis S. COVID-19: a comparison of time series methods to forecast percentage of active cases per population. Appl Sci. 2020;10(11):3880. doi:10.3390/app10113880
4. Tiwari D, Bhati BS, Al-Turjman F, Nagpal B. Pandemic coronavirus disease (Covid-19): world effects analysis and prediction using machine-learning techniques. Expert Syst. 2021;39(3):e12714. doi:10.1111/esy.12714
5. Ghazaly NM, Abdel-Fattah MA, El-Aziz AAA. Novel coronavirus forecasting model using nonlinear autoregressive artificial neural network. Int J Adv Sci Technol. 2020;29(5):1831-1849.
6. Ratre S, Sharma R, Kevadia M, Bathina V. COVID-19 prediction using regression; 2021.
7. Kumar P, Kalita H, Patiaryya S, et al. Forecasting the dynamics of COVID-19 pandemic in top 15 countries in April 2020: ARIMA model with machine learning approach. bioRxiv; 2020. doi:10.1101/2020.03.30.0046227
8. Tuli S, Tuli S, Tuli R, Gill SS. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Intern Things. 2020;11(100222):1-100222. doi:10.1101/j.imatr.2020.100222
9. Gothai E, Thamilselvan R, Rajalaxmi RR, Sadana RM, Ragavi A, Sakthivel R. Prediction of covid-19 growth and trend using machine learning approach. Mater Today. 2021. doi:10.1016/j.matpr.2021.04.051
10. Saxen K, Vohra A, Rana A. Covid-19 Outbreak Prediction Through Prophet Base Model.
11. Rauf HT, Lali MIU, Khan MA, et al. Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks. Pers Ubiquit Comput. 2021;25:1-18. doi:10.1007/s00779-020-01494-0
12. Malik Z, Atlam ES, Hassanien AE, Dagniew G, Elhosseini MA, Gad I. Association between weather data and COVID-19 pandemic predicting mortality rate: machine learning approaches. Chaos Solit Fract. 2020;138:110137. doi:10.1016/j.chaos.2020.110137
13. Malik Z, Atlam ES, Ewis A, et al. The COVID-19 pandemic: prediction study based on machine learning models. Environ Sci Pollut Res Int. 2021;28(30):40496-40506. doi:10.1007/s11356-021-13824-7
14. Ardabili S, Mosavi A, Band SS, Varkonyi-Koczy AR. Coronavirus disease (COVID-19) global prediction using hybrid artificial intelligence method of ANN trained with grey wolf optimizer. In: 2020 IEEE 3rd International Conference and Workshop in Obuda on Electrical and Power Engineering (CANDO-EPE). IEEE; 2020.
15. Khakharia A, Shah V, Jain S, et al. Outbreak prediction of COVID-19 for dense and populated countries using machine learning. Ann Data Sci. 2021;8(1):1-19. doi:10.1007/s40745-020-00314-9
16. Alizadehsani R, Sharifzadeh D, Izadi NH, et al. Uncertainty-aware semi-supervised method using large unlabeled and limited labeled COVID-19 data. ACM Trans Multimed Comput Commun Appl. 2021;17(3s):1-24. doi:10.1145/3462635
17. Dash S, Chakraborty C, Giri SK, Pani SK. Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics. Pattern Recogn Lett. 2021;151:69-75. doi:10.1016/j.patrec.2021.07.027
18. Asgharnezhad H, Shamsi A, Alizadehsani R, et al. Objective evaluation of deep uncertainty predictions for COVID-19 detection. Sci Rep. 2022;12(1):815. doi:10.1038/s41598-020-22562-6
19. Sujath R, Chatterjee JM, Hassanien AE. A machine learning forecasting model for COVID-19 pandemic in India. Stoch Environ Res Risk Assess. 2020;34(7):959-972. doi:10.1007/s00477-020-01827-8
20. Joloudari JH, Azizi F, Nodehi I, et al. DNN-GFE: A Deep Neural Network Model Combined with Global Feature Extractor for COVID-19 Diagnosis Based on CT Scan Images. 2021;6330.
21. Hyndman RJ, Koehler AB, Snyder RD, Grose S. A state-space framework for automatic forecasting using exponential smoothing methods. Int J Forecast. 2002;18(3):439-454.
22. Ayobi N, Sharifzadeh D, Alizadehsani R, et al. Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods. Results Phys. 2021;27(104495).104495. doi:10.1016/j.rinp.2021.104495
23. Khoezimeh F, Sharifzadeh D, Izadi NH, et al. Combining a convolutional neural network with autoencoders to predict the survival chance of COVID-19 patients. Sci Rep. 2021;11(1):15343. doi:10.1038/s41598-021-93543-8
24. Omran NF, Ghany SF, Saleh H, Ali AA, Gumaei A, Al-Rakhami M. Applying deep learning methods on time-series data for forecasting COVID-19 in Egypt, Kuwait, and Saudi Arabia. Complexity. 2021;2021(7):1-13.
25. Sharifzadeh D, Alizadehsani R, Roshanzamir M, et al. Fusion of convolution neural network, support vector machine and Sobel filter for accurate detection of COVID-19 patients using X-ray images. Biomed Signal Process Control. 2021;68(102622):102622. doi:10.1016/j.bspc.2021.102622
26. BBC News. Coronavirus: Up to 70% of Germany could become infected—Merkel. Accessed March 15, 2020. https://www.bbc.co.uk/news/world-europe germany-51835856
27. World Health Organisation (WHO): https://www.who.int/emergencies/diseases/novel-coronavirus-2019
28. Centers for Disease Control and Prevention: https://www.cdc.gov/coronavirus/2019-ncov

How to cite this article: Jain P, Sahu S. Prediction and forecasting of worldwide corona virus (COVID-19) outbreak using time series and machine learning. Concurrency Computat Pract Exp. 2022;34(26):e7286. doi: 10.1002/cpe.7286