Pertanika J. Sci. & Technol. 31 (4): 1991 - 2011 (2023)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Forecasting Geo Location of COVID-19 Herd

Divyansh Agarwal[1], Nishita Patnaik[1], Aravind Harinarayanan[1], Sudha Senthilkumar[1]*, Brindha Krishnamurthy[2] and Kathiravan Srinivasan[1]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India
[2]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India

ABSTRACT

Thanks to the growth in data storage capacity, nowadays, researchers can use years' worth of mathematical models and depend on past datasets. A pattern of all pandemics can be identified through the assistance of Machine Learning. The movement of the COVID-19 herd and any future pandemic can be predicted. These predictions will vary based on the dataset, but it will allow the preparation beforehand and stop the spreading of COVID-19. This study focuses on developing Spatio-temporal models using Machine Learning to produce a predictive visualized heat regional map of COVID-19 worldwide. Different models of Machine Learning are compared using John Hopkins University dataset. This study has compared well-known basic models like Support Vector Machine (SVM), Prophet, Bayesian Ridge Regression, and Polynomial Regression. Based on the comparison of various metrics of the Support Vector Machine, Polynomial Regression Model was found to be better and hence can be assumed to give good results for long-term prediction. On the other hand, ARIMA, Prophet Model, and Bayesian Ridge Reduction models are good for short-term predictions. The metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE) are better for Support Vector Machines compared to other models. The metrics such as R^2 Score and Adjusted R-Square are better for the polynomial Regression model.

Keywords: ARIMA, Bayesian ridge regression, COVID-19, polynomial regression, predictions, prophet, support vector machine (SVM)

ARTICLE INFO

Article history:
Received: 16 July 2022
Accepted: 02 November 2022
Published: 13 June 2023

DOI: https://doi.org/10.47836/pjst.31.4.23

E-mail addresses:
divyansh.agarwal2019@vitstudent.ac.in (Divyansh Agarwal)
nishita.patnaik2019@vitstudent.ac.in (Nishita Patnaik)
inaravind.harinarayanan2019@vitstudent.ac.in (Aravind Harinarayanan)
sudha.s@vit.ac.in (Sudha Senthilkumar)
brindha.k@vit.ac.in (Brindha Krishnamurthy)
kathiravan.srinivasan@vit.ac.in (Kathiravan Srinivasan)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Over the last decade, Machine Learning (ML) has emerged as a major field of study as it has solved many complex real-world problems. The forecasting areas are the main important areas of ML. Various ML algorithms forecast future events for applications like weather, disease, and stock market prediction. ML techniques have been used to forecast diseases such as breast cancer, coronary artery, and cardiovascular disease. Specifically, this article aims to provide real forecasting of COVID-19 cases.

Five types of pandemics have been seen in recent decades: H1N1 (2009–2014), Polio (2014–r2016), Zika (2016–2019), and Ebola (Democratic Republic of Congo). On January 30, 2020, the World Health Organization registered COVID-19 as the sixth worldwide pandemic. Deaths and morbidities cause a high and huge amount of economic loss worldwide. Almost every country infected with a pandemic experienced serious health-related concerns, and pandemics negatively impact the socioeconomic situation. The COVID-19 pandemic is the greatest threat to public health. WHO has received 608,328,548 reported cases worldwide, with 6,501,469 deaths (Li et al., 2020). As of September 16, 2022, Table 1 shows the COVID-19-affected countries worldwide.

Currently, computerized data are collected in a way that makes it difficult to analyze and predict disease growth locally and globally. The disease and its progression can be successfully mapped using ML algorithms. In order to train an ML model to predict the number of global confirmed cases prone to the disease in the coming days, supervised models with associated algorithms (like LR, SVR, and time series algorithms) are used to analyze data for regression and classification. This proposal collects and pre-processes the global dataset and extracts the number of confirmed cases up to a particular date, serving as the model's training set. In order to predict the growth of cases in the upcoming days, the model is being trained by supervised ML algorithms.

Table 1
COVID-19 affected countries worldwide as of September 16, 2022 ("World Health Organization," 2020)

Country name	Total cases	Total deaths	Active cases reported in last 7 days
USA	94,237,260	1,041,323	414,468
India	44,522,777	528,273	38,048
France	33,796,693	151,089	161,584
Australia	10,153,910	50,077	14,682
South Africa	4,015,347	102,146	1,626
UK	23,585,309	189,484	26,045
Brazil	**34,558,902**	**685,121**	**59,079**

Based on massive datasets, we aim to produce an ML model that aims at high-precision prediction of the movement of COVID-19 cases (He et al., 2021). This study will help prepare before the peak/waves of COVID-19 strikes. This model's data collection can be used for other (Liu et al., 2018) pandemics over future centuries and allows for stopping/predicting/simulating viruses and their effect. With the help of various parameters, features of regional heat maps changing with time can be added (Shilo et al., 2020), which gives any person easy understandability (Looi, 2020). With the rising number of vaccinations for COVID-19, the number of cases and deaths might change unpredictably.

The data was taken for one year only as there were chances of oversampling of data (Pullano et al., 2020). This study seeks to exhibit future forecasting on the overall count of total cases in India in the next 30 days to contribute to controlling the spread and growing count of current cases in India. Models were tested with the last month of data. For every date, the global map prediction was analyzed and presented as a map, with the impact factor of each country (Allwood et al., 2022; Garrido et al., 2022; Mogensen et al., 2022).

The model is assessed for performance measures like accuracy during this procedure. Accuracy calculates the model's performance over unknown data by splitting the number of adequately predicted features by the number of accessible features to be forecasted. Many ML methods are applied to anticipate and forecast future appearances. SVM classifier, Prophet, Bayesian Networks, Polynomial Ridge Regression, and ARIMA are the ML methods used for prediction and analysis. According to the information conducted for this study, Support Vector Machine and Polynomial regression models were found to be better and hence can be assumed to give good results for long-term prediction. They had the highest accuracy of all classifiers, and we tested prediction methods when evaluating model efficiency (Bird et al., 2020).

Particularly, our research focused on the following objectives.

- The various models are chosen to compare the results (Lu et al., 2020). The three models: SVM, Polynomial Regression, and Bayesian Ridge Model, use over one year of data to get trained and produce the result for COVID-19 cases (Aarthi et al., 2018; Klyushin, 2020).
- We have added a visualization map of the data collected for over two years, showcasing the (Mudenda et al., 2021) region-wise cases, deaths, and recoveries with increasing dates. We have used choropleth maps to implement this visualization feature to understand the disease with increasing time better. It showcases the impact factor for each country with the help of a color bar which helps to analyze which country is impacted severely by COVID-19 immediately.
- The ML models were trained using the COVID-19 Johns Hopkins dataset. This dataset is pre-processed and segregated into three subsets: training dataset (70%) and, validation dataset (15%), testing dataset (15%). The model performance is

evaluated using key metrics such as R^2 score, Adjusted R^2 score, mean squared error (MSE), mean absolute error (MAE), and root mean square error (RMSE). We presented the comparison of performance metrics with various models.

RELATED WORKS

Bae et al. (2021) proposed that SEIR and Regression models are used to anticipate disease infection worldwide. The number of reported cases over the following 21 days was anticipated based on 62 days of training and 5 days of testing. The statistics were not steady and exhibited an exponential rise after 40 days, starting on January 22, 2020. Overfitting is still a serious issue in disease transmission (Bae et al., 2021).

Rustam et al. (2020) suggested that the model uses four standard forecasting models: SVM, LASSO, LR, and ES. Each model estimates three data types for the next 10 days, along with the number of freshly infected persons, fatalities, and recoveries. The findings demonstrate that the ES beats all other models closely by the LR and LASSO (Rustam et al., 2020).

Liu et al. (2020) discussed that the clustering strategy, and even a data augmentation technique, is used in machine learning methodology to leverage the geographic synchronicity of COVID-19 movement across Chinese regions. Their model can produce steady forecasts two days before the actual event. The limitation of this model is that there was a constant drop in COVID-19 cases reported over the testing period of our strategies; therefore, this approach could not be tested for its ability to recognize pandemic spikes across regions (Liu et al.,2020).

Wynants et al. (2020) suggested that SEIR and Regression models are used to anticipate disease infection worldwide. The number of reported cases over the following 21 days was anticipated based on 62 days of training and 5 days of testing. The statistics were not steady and exhibited an exponential rise after 40 days, starting on January 22, 2020. Overfitting is still a serious issue in disease transmission time series data (Wynants et al., 2020).

Mahdavi et al. (2021) presented that the model uses four standard forecasting models: SVM, LASSO, LR, and ES. Each model estimates three data types for the next 10 days, along with the number of freshly infected persons, fatalities, and recoveries. The findings demonstrate that the ES beats all other models, closely by the LR and LASSO, while the SVM performs poorly in all prediction situations given the available data (Mahdavi et al., 2021; Quah et al., 2020).

Pan et al. (2021) developed the spatial-temporal-susceptible-infected-removed model (STSIR). To make SIR a dynamic system, they combine both intra-city and inter-city mobility indices. This model can accurately predict the total pandemic scale using the observable index. In terms of predicting the final scale of the pandemic, this model achieves an MAE of 7.76 (Pan et al., 2021).

With Olszewski et al. (2012) model, it is possible to analyze changes in the number of cases over time and in space. In addition to considering spatial conditions in terms of population distributions, such as places of work, rest, and residence, the methodology also uses multi-agent modeling to examine spatial interactions. The model was further enhanced by introducing a multi-variant vaccination policy (Olszewski et al., 2021).

SVR and ARIMA models were compared to predict daily imported new COVID-19 cases in Shanghai, China, by Zhao and Zhang (2022). Their epidemic trend was predicted using ARIMA models. The SVR model outperformed the ARIMA model in the study. An early warning of the COVID-19 outbreak can help prevent and control the outbreak in Shanghai, China.

PRELIMINARIES

Machine Learning Models

SVM Model. A Supervised Learning model is used for solving classification and regression problems. Using SVM, n-dimensional space can be easily categorized into classes so that future data points can be easily assigned to the right category. (Shaukat et al., 2015)

The Support Vector Machine (SVM) is a powerful learning method for binary classification. Its main purpose isto find the best hyperplane for correctly dividing data into two groups. Aside from that, it excels at dealing with high-dimensional and nonlinear data. As a result, we used the SVM model to anticipate confirmed instances, as shown in Equation 1.

$$\boldsymbol{f(x) = w^T x + b = \sum_{j=1}^{M} w_j x_j \ + b = 0} \tag{1}$$

Polynomial Regression Model. Polynomial Regression is a method that handles an nth-degree polynomial to characterize the link be- tween a dependent variable (y) and an independent variable (x). Equation 2 is as follows:

$$\boldsymbol{y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3 + \cdots b_n x_1^n} \tag{2}$$

It is also known as the special case of Multiple Linear Regression in machine learning because we modify the Multiple Linear Regression equation with some polynomial terms to make it polynomial regression.

Prophet Model. The Prophet library is a free, open-source library for forecasting univariate time series datasets. It is easy to useand designed to automatically find a good set of hyperparametersfor the model to generate reliable data forecasts. The prophet forecasting model employs a perishable model comprising three primary factors:trends, holidays, and seasonality. They are shown in the following Equation 3.

$$\boldsymbol{y(t) = g(t) + s(t) + h(t)} + \in \tag{3}$$

g(t) is a section-wise logistical or linear growth curve usedin statistics to describe non-periodic variations,
s(t) is seasonal fluctuations that can occur weekly oryearly,
h(t) is the effects of user-provided holidays with varying schedules,
ε estimates of error terms used in any major changes not included in the model.

Exponential smoothing and a prophet use the same method to simulate periodicity as a supplement component. Because exponentially declining weights areallocated owing to the appearance of the data, the dominant facts are given equal importance in predicting than the earlier observations in exponential smoothing.

ARIMA. ARIMA (Auto Regressive Integrated Moving Average) is a type of model that describes a time series based on its prior values, i.e., lags and delayed prediction errors, so that it can be used to forecast future values. ARIMA models can model any non-seasonal time series with patterns, not random white noise.

An ARIMA model is defined by three terms: p, d, and q, where p is the ARIMA order and q is the order of the Moving Average (MA). The term d denotes the number of differencing necessary to make the time series steady. As a result, the generic ARIMA Equation 4 is shown below:

$$\boldsymbol{Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \ + \ \varepsilon_{\ t}} \tag{4}$$

where Y_t is the anticipated target is a constant value plus a linear combination of Y delays taken till p lags andthe mixture of delayed prediction error taken till q lags.

Similarly, a pure Moving Average (MA alone) model is one in which Y_t is determined only by the delayed forecast errors, as shown in Equation 5:

$$\boldsymbol{Y_t = \alpha + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} \ldots + \phi_q \varepsilon_{t-q}} \tag{5}$$

where the variance of the error term is the auto-regressive model errors for the relevant delays, the errors t and t-1 represent the results of Equations 6 and 7.

$$\boldsymbol{Y_t = Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_0 Y_0 \ + \ \varepsilon_t} \tag{6}$$

$$\boldsymbol{Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} \ldots + \beta_0 Y_0 \ + \ \varepsilon_{t-1}} \tag{7}$$

In an ARIMA model, the time series is divided into two classes at most once to keep it stationary, and the AR and MA conditions are mingled. As a result, Equation 8 is shown below:

$$\boldsymbol{Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p}\ \varepsilon_{\,t} + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} \ldots + \phi_q \varepsilon_{t-q}} \tag{8}$$

Bayesian Ridge Polynomial Regression Model. Bayesian regression defends poorly distributed data using natural processes.Instead of being estimated as a single number, the outputor answer 'y' is designed to select from the probability distribution.
The answer y is assumed to be a Gaussian distribution with X_w to design a fully probabilistic model, as shown below in Equation 9.

$$\boldsymbol{p(y|X, w, \propto) = N(y|X_w, \propto)} \tag{9}$$

Bayesian Ridge regression, which computes a probabilistic model of regression issues, is one of the most useful variants of Bayesian regression. Spherical Gaussian provides the following antecedent for the coefficient w shown in Equation 10.

$$\boldsymbol{p\,(w\,|\lambda\,) = N\,(w\,|0, \lambda^{-1} I_p)} \tag{10}$$

Gamma distributions, the corresponding prior for Gaussian accuracy, are chosen as throughout the model fit, the regularization parameters, and the parameters w, alpha, and gamma, are computed simultaneously, with the log residual likelihood being maximized. Hyper-parameters such as lambda and alpha were tweaked to generate considerably better outcomes.

DATASET DETAILS

The Johns Hopkins University COVID-19 Data Repository is referred for this study and used by the Center for Systems Science and Engineering (CSSE). The JHU CSSE's GitHub public repository can be used for educational and academic research. It is publicly available at the URL https://github.com/CSSEGISandData/COVID-19. csse_covid_19_time_series data consists of two cumulative datasets that Johns Hopkins produces, with the latest files updated daily at 23:59 UTC. The attributes such as Country, State/City, Confirmed cases, Recovered cases, and Death counts were considered for evaluation. The dataset contains 192445 records with 10 attributes, 134711 records for training, 28867 records for testing, and 28867 records are used for validation.

EVALUATION PARAMETERS

The parameters Root Mean Squared Error (RMSE), R^2 Score, Mean square Error (MSE), Mean absolute error (MAE), and Adjusted R-Square (R^2adjusted) are considered in the evaluation and comparison of Models.

Mean Square Error (MSE)

The square error suggests a different technique for evaluating the existence of reviewing models. MSE takes the information distance, concentrates on the return line, and squares it. Coupling is essential because it removes the minus number and adds weight to the primary differentiation. A little square error indicates how near you are to obtaining the optimal condition line.

It is calculated by finding the variance among the original and predicted values of the data, then calculating the square of this result, and then taking the average as shown in Equation 11.

$$\boldsymbol{MSE = \frac{1}{N}(\sum_{J=1}^{n}(act_{value} - pred_{value})} \tag{11}$$

N is the total number of observations, act_value indicates the actual value, and pred_value indicates the predicted value.

Root Mean Squared Error (RMSE)

When a prediction is made on a dataset, the RMSE is defined as the standard deviation of errors. It is similar to MSE; however, the root of its value is considered, as shown in Equation 12.

$$\boldsymbol{RMSE = \sqrt{MSE}} \tag{12}$$

R^2 Score

The R-squared (R^2) scale is a calculating scale used to evaluate review models' presentation. The degree of the relationship between adaptation and inversion models is determined by a positive size ranging from 0 to 100 percent (Equation 13).

$$\boldsymbol{R^2 = 1 - \frac{RSS}{TSS}} \tag{13}$$

RSS denotes squares of residuals sum, and TSS denotes the total sum of squares.

Adjusted R^2 Score

It is a variant of the R^2 score. The adjusted R^2 score automatically adjusts according to various features considered for the prediction model. This metric value increases in case new features lead to improvement and decreases in case the new features do not lead to much improvement. Equation 14 denotes the formula for Adjusted R^2 score calculation.

$$\boldsymbol{Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{(N - p - 1)}} \tag{14}$$

R^2 denotes sample R-square, N is the total sample size, and p is the predicators count.

Mean Absolute Error (MAE)

The average of all individual differences between the model predictions and actual data is computed by averaging test data (Equation 15).

$$\boldsymbol{MAE = \frac{\sum_{j=1}^{m} \left| y_j - x_j \right|}{m}} \tag{15}$$

Y_j indicates the prediction value, x_j indicates the true value, and m indicates the total number of data samples.

METHODOLOGY

Many ML methods are applied in this study to anticipate and forecast future appearances of COVID-19 cases, such as SVM classifier, Prophet, Bayesian Networks, Polynomial Ridge Regression, and ARIMA. The model having the highest accuracy is picked for forecasting or prediction during the model evaluation. This proposed work aims to achieve future forecasting for the next 10 days to identify the total number of recovery cases, daily confirmed cases and death rate.

The dataset is divided into 70% for training, 15% for testing, and 15% for validation. The six models, SVM, Polynomial Regression, Bayesian Ridge Model, ARIMA, and Prophet, use data over a range of 1-year data to get trained, validated, and tested to produce results with better precision and accuracy.

First, the data set is retrieved from a webpage. Web scraping makes the extraction feasible (Rustam et al., 2020). Following the conclusion of web scraping, the data set is related to data wrangle and pre-processing before being saved to the local storage. The pre-processed data is now shown for a flawless result, a high-level data set summary. The dataset is divided into 70% for training, 15% for testing, and 15% for validation. The 15% of test data collected from the same dataset is kept hidden from the model during one of the training periods; hiding a portion of the dataset aids in discovery. Determine if the model has under-fit or over-fit and some of the most significant challenges while developing any model. The ARIMA Model is trained by supplying a training dataset. The model is ready for testing once it has been trained. One of the most difficult aspects of training any model is over-fitting. Before the model is tested, the Facebook Prophet (Arabi et al., 2020) provides the dates for the next 30 days as well as the timestamp to forecast. Finally, the time and date stamp provided by Facebook Prophet is applied to the test data set. The model is currently being trained on the entire number of active case patterns. The ARIMA model has been analyzed and reported on key metrics such as R-Squared Score, MAE, MSE, and RMSE (Grasselli et al., 2020)

The Prophet and ARIMA are specifically designed mostly for prediction. Hence, we have trained these models with data over 2 years. Bayesian Ridge polynomial reduction, Prophet, SVM, and Polynomial regression give almost similar results. These various models are trained on the confirmed, recovered, and death cases. Further, ML models have been evaluated using various key metrics such as MSE, RMSE, R 2 -score, R 2 adjusted score, and MAE, and results are produced.

The system architecture diagram for data processing, model development, and deployment, followed by the monitoring, is shown in Figure 1.

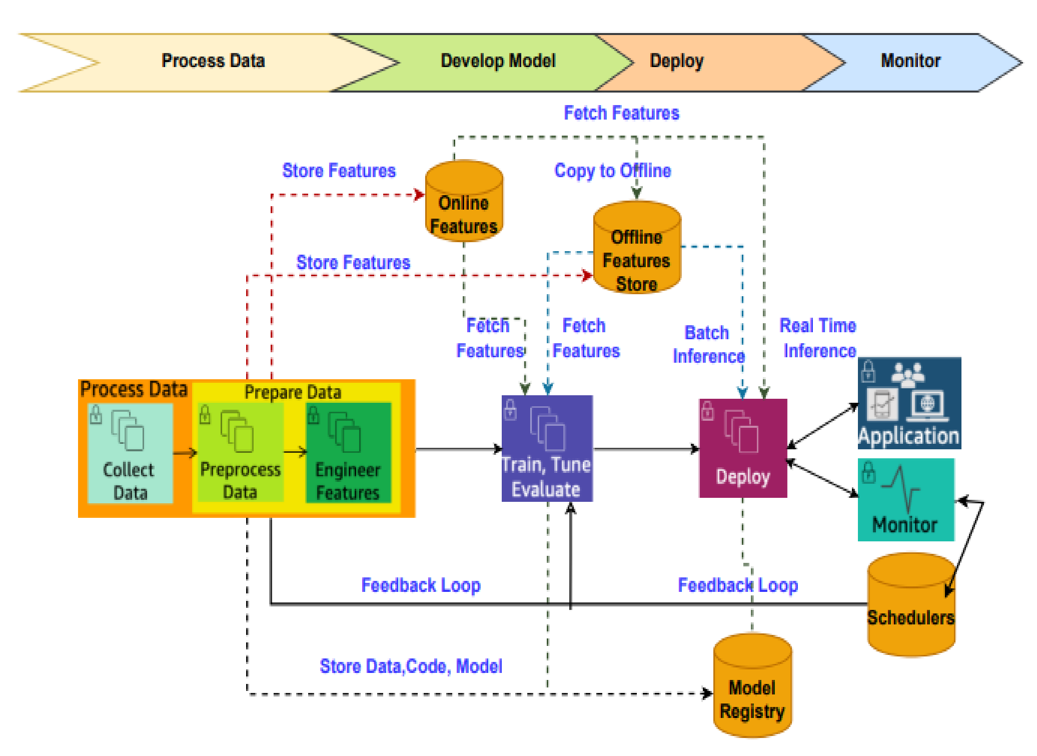


Figure 1. Proposed system architecture

RESULT

To analyze the result performance of all the models tested with three different sizes of testing, training, and validation datasets. Initially, the dataset is divided into 70% for training, 15% for testing, and 15% for validation. This result performance is shown in Table 2, with various metrics evaluated in Test 1. Again, the same dataset is split into 80% for training, 10% for testing, and 10% for validation. This result is indicated in Table 3, with various metrics evaluated in Test 2. Finally, the same dataset is divided into 76% for training, 12% for testing, and 12% for validation. This experimental result is indicated in Table 4, with various metrics evaluated in Test 3. Among all the tests Support Vector Machine and Polynomial regression models were found to be better and hence can be assumed to give good results for long-term prediction.

SVM Model

The SVM Prediction and Test Data are very close, showing good results with large datasets (Abedini et al., 2017). It takes past data and does the analysis based on that. A higher percentage of training will produce more accurate results. There are chances of overfitting with large amounts of data, which is avoided by doing the pre-processing. Figure 2 shows the SVM model predictions. Big-O notation is used to calculate the complexity of an algorithm. An algorithm will process amounts of data, where N is a symbol for amounts of data. The lower bound, upper bound and total number of loop iterations determine the complexity. The computational complexity of SVM is O(n^3).

Polynomial Regression

The Polynomial Regression model (Bae et al., 2021) is much closer to the test data compared to the SVM Model, so it shows that we can use this model for prediction. This model also provides good results when trained with a large dataset. Polynomial Regression also takes the past trend and forecasts the COVID-19 cases. In this model, factors affect it largely. So,

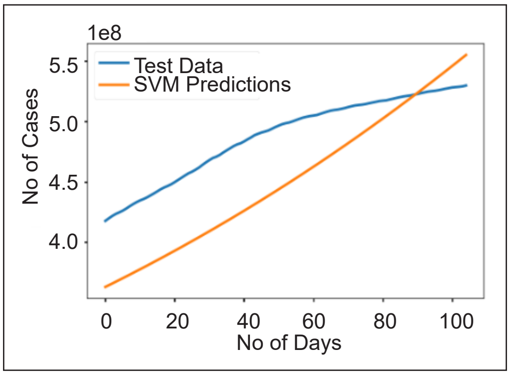


Figure 2. SVM model predictions

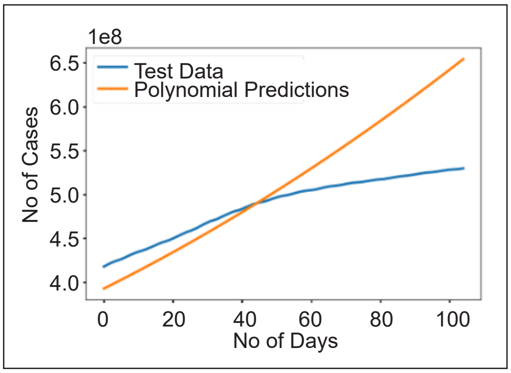


Figure 3. Polynomial regression predictions

it becomes necessary to take appropriate parameters for better prediction. Figure 3 denotes the polynomial regression predictions. It uses a linear regression model to fit complex, nonlinear functions and datasets. In the m degree polynomial regression with n measured values ($n \geq m + 1$), the computational complexity is $O(n^2m)$.

Prophet

The Prophet also gives a good result but not as good as polynomial regression. It is better than the SVM model to be used for prediction. The prophet prediction is slightly lower than the test data. Figure 4 shows the Prophet Facebook time series model. It takes all the past trends and forecasts the trend. It is good for a large dataset but cannot be forecast longer. It gives good results when it is forecast within 25 days.

ARIMA

ARIMA is giving a good result with almost overlapping the test data, but later as the COVID-19 cases dip, it cannot continue. This model is specifically built for time series models. It gives good results for small values but also needs to be checked with large ones. Figure 5 shows that ARIMA gives a good result when the forecast time is within 25 days. The model does not face any over-fitting problems, or it can neglect some noise in the data. The computational complexity of the ARIMA model is $O(n^2T)$, where n is the number of parameters ($n = p+q+P+Q$) where p, q & P, and Q are non-seasonal and seasonal orders and T is the length of the time series.

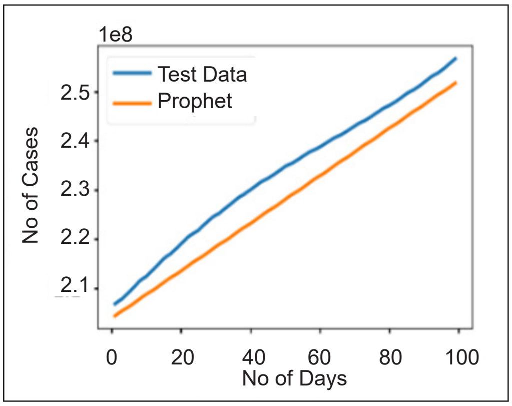


Figure 4. Prophet Facebook time series model

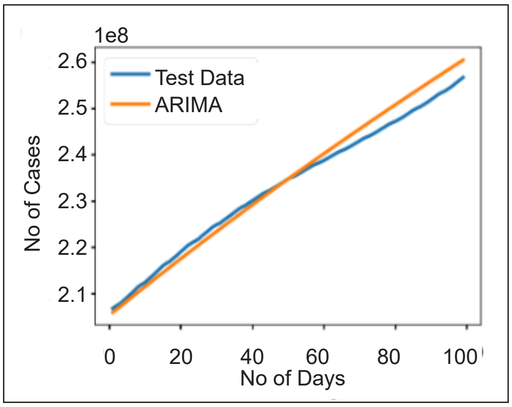


Figure 5. ARIMA model predictions

Bayesian Ridge Polynomial Reduction

Bayesian Ridge is almost giving the same result as SVM. It is not giving good results with test data. Figure 6 shows the Bayesian Ridge Polynomial Reduction Model. The model ispurely based on the number of days; hence noise should be avoided, or it can lead to over-fitting. The trained models can be tested with all the data and also predict for the forecast 10 days. This model was tested with the real-time dataset from March 2020 to now and forecasted for the next 10 days. The over-fitting problem can be solved by removing noise during a pre-processing phase. Linear Regression has a runtime complexity of O(k). Linear regression has a long training time but is very efficient during testing. Tests/predictions are performed in O(k) time, where k is the data's number of features/dimensions.

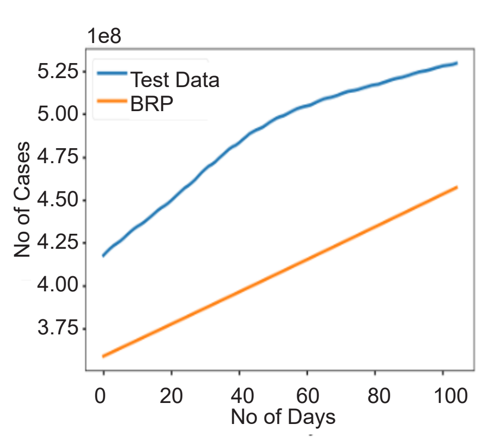


Figure 6. Bayesian Ridge Polynomial Reduction Prediction Model

Growth Factor

It is not possible to determine the growth factor from the data since we only have access to the number of cases per day. An experimental daily observation can be used to determine the growth factor with various statistical models. The average growth factor is calculated based on the 500 days of data observations. The performance was tested with large datasets. It is the simplest method for forecasting trends. It calculates the average growth of COVID-19 cases daily based on the prediction. Figure 7 denotes the prediction growth factor.

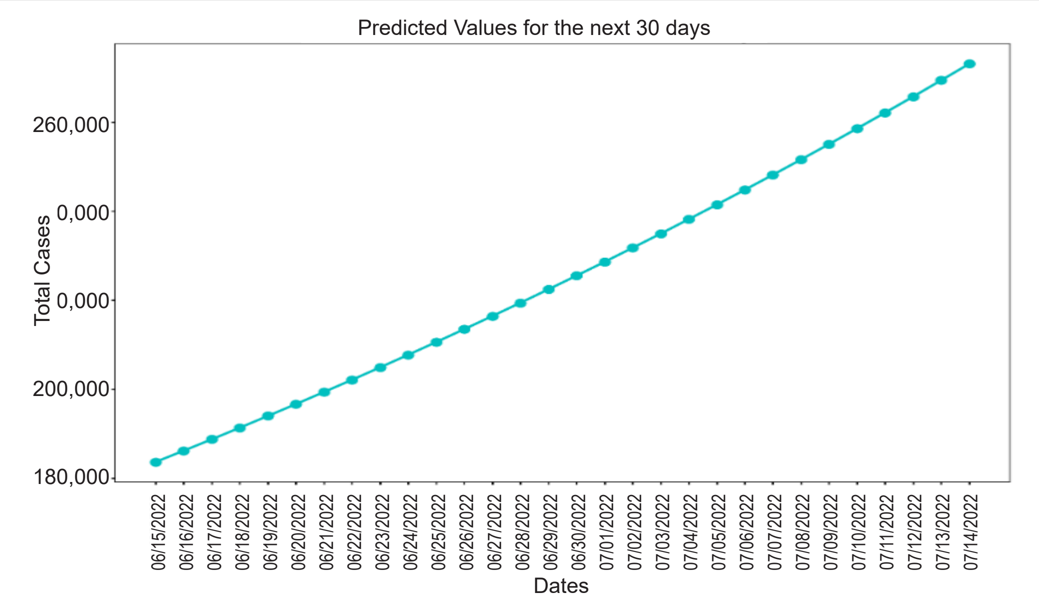


Figure 7. Prediction growth factor

SVM Prediction

Figure 8 shows the SVM Model predicts that for the next 500 days from January 2022. The SVM Model produces an accurate result for past trends.

Polynomial Regression

This model predicts good results for large datasets. The growth factor is considered for monthly or yearly COVID-19 cases. Figure 9 shows Polynomial Regression based on growth factor with minimum deviation. This model produces a better result for the present trend.

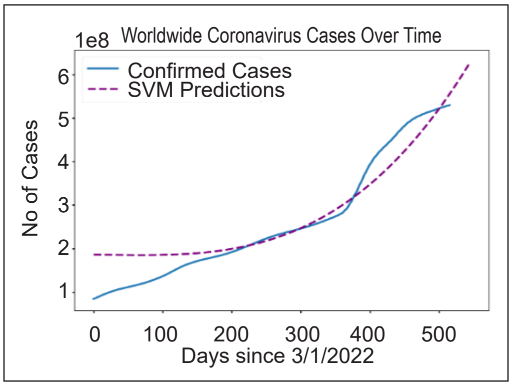


Figure 8. SVM Prediction based on growth factor

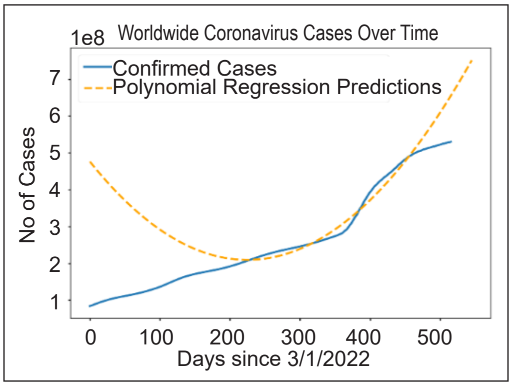


Figure 9. Polynomial Regression based on growth factor

Bayesian Ridge Regression

This model predicts better results for large datasets. The growth factor is considered for monthly or yearly COVID-19 cases. Figure 10 shows Bayesian Ridge Regression based on growth factor with minimum deviation.

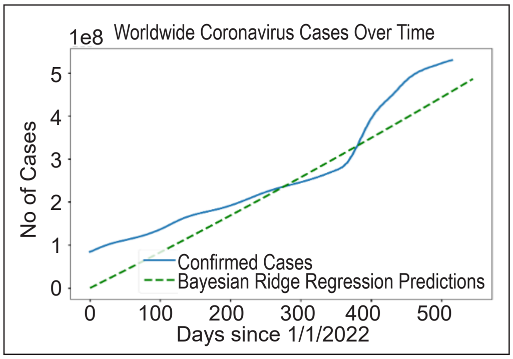


Figure 10. Bayesian Ridge Regression based on growth factor

ARIMA Prediction

It predicts and forecasts the trends based on the input data. It gives the best result for past and future trends. Figure 11 denotes that the dotted yellow line predicts the COVID-19 cases compared with actual data. This model produces the best results for a small period of forecast days.

Prophet Prediction

The Prophet Model is the open-source time series forecasting algorithm developed by Facebook. The black dots that form a line in the figure represent historical training data in the model. The blue line represents the forecast or fitted curve generated for the past and the future. The light blue shaded region represents the brands of uncertainty. In COVID-19, the growth factor considered for the number of days was 19 cases. Based on the forecasting days, it gives peak results for 10-15 days. Figure 12 depicts the Facebook Prophet model based on the growth factor (Estenssoro et al., 2022). The graph shows the predicted number of cases as well as the uncertainty band that it provides based on seasonality.

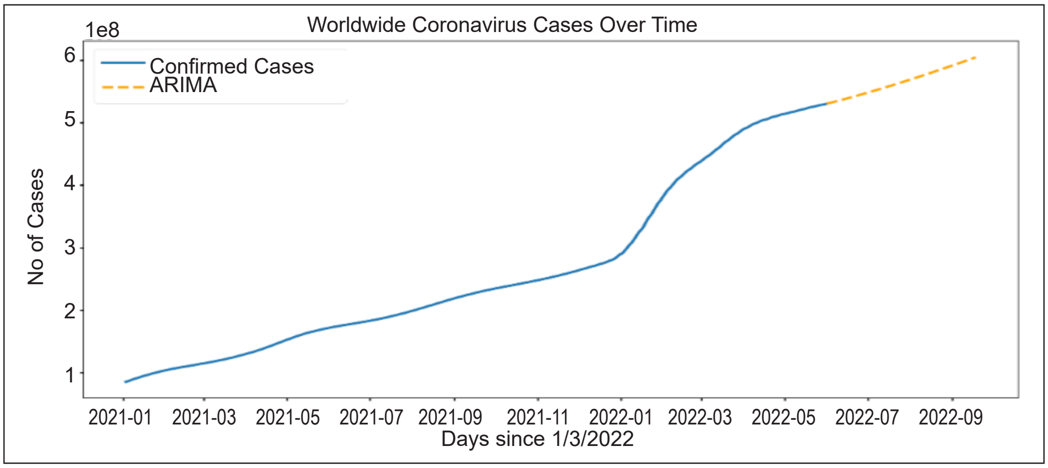


Figure 11. ARIMA prediction based on growth factor

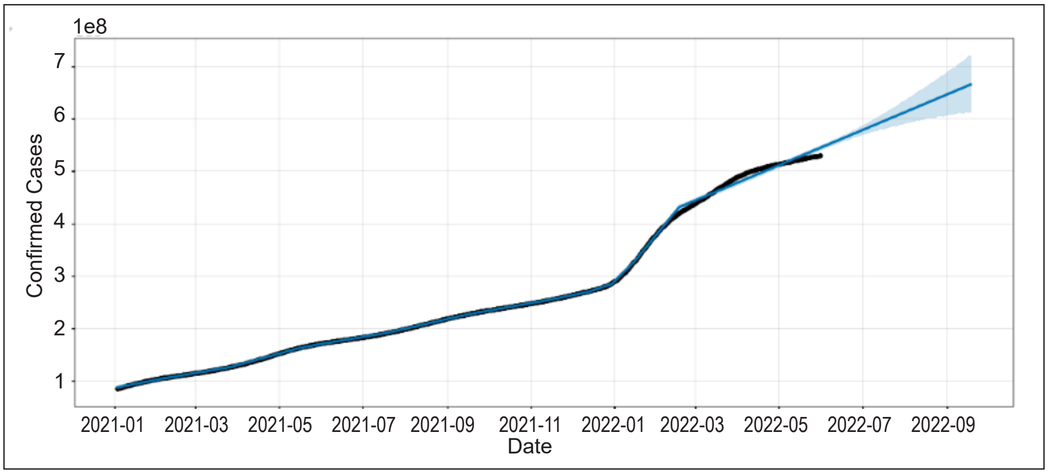
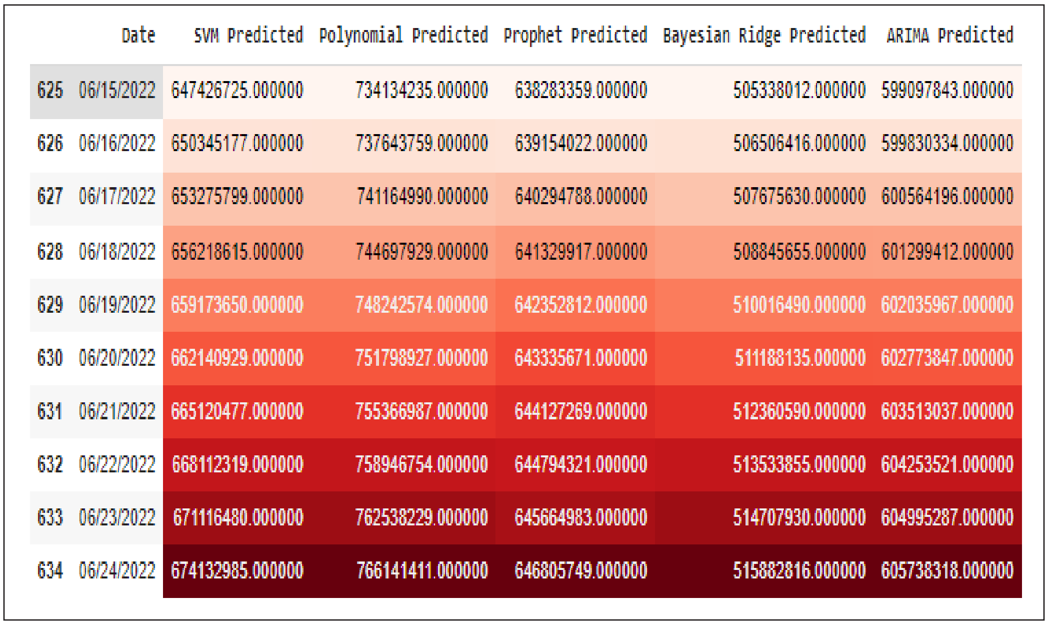


Figure 12. Facebook Prophet model based on growth factor

DISCUSSION

The Polynomial Regression and Prophet model produces good results with the number of cases it has predicted compared to other models we considered for the COVID-19 prediction. The model has predicted the cases from June 15, to June 24, 2022, in which prophet and polynomial produce good results. The result in Figure 13 denotes that the SVM model is good for a stable increase in COVID-19 cases for long-term prediction. However, given different scenarios, ARIMA and Prophet are good for predicting the next 5–10 days. Figure 13 shows the Heat Maps prediction with various models used in this proposed work (Dong et al., 2020).



	Date	SVM Predicted	Polynomial Predicted	Prophet Predicted	Bayesian Ridge Predicted	ARIMA Predicted
625	06/15/2022	647426725.000000	734134235.000000	638283359.000000	505338012.000000	599097843.000000
626	06/16/2022	650345177.000000	737643759.000000	639154022.000000	506506416.000000	599830334.000000
627	06/17/2022	653275799.000000	741164990.000000	640294788.000000	507675630.000000	600564196.000000
628	06/18/2022	656218615.000000	744697929.000000	641329917.000000	508845655.000000	601299412.000000
629	06/19/2022	659173650.000000	748242574.000000	642352812.000000	510016490.000000	602035967.000000
630	06/20/2022	662140929.000000	751798927.000000	643335671.000000	511188135.000000	602773847.000000
631	06/21/2022	665120477.000000	755366987.000000	644127269.000000	512360590.000000	603513037.000000
632	06/22/2022	668112319.000000	758946754.000000	644794321.000000	513533855.000000	604253521.000000
633	06/23/2022	671116480.000000	762538229.000000	645664983.000000	514707930.000000	604995287.000000
634	06/24/2022	674132985.000000	766141411.000000	646805749.000000	515882816.000000	605738318.000000

Figure 13. Heat Maps prediction with various models

Global Visualization

Figure 14 showcases the data visualization, the trends of COVID-19 for each country, and the number of cases, deaths, and recovery for each day (27–29). It provides the best and easiest manner to understandand analyze the pandemic (Franch-Pardo et al., 2020). The ranges are denoted in different shades, starting from 0 cases to 1–5,000, 5,001–50,000, 50,001–5,000,000, and > 5,000,000.

Tables 2, 3 and 4 compare performance metrics with machine learning models for different test data sets. The dataset is divided into 70% for training, 15% for testing, and 15% for validation in Test 1, and the performance of various evaluation metrics are shown in Table 2. Further, the same dataset is split into 80% for training, 10% for testing, and 10% for validation in Test 2, and the performance result for various metrics is indicated in Table 3. Finally, the same dataset is divided into 76% for training, 12% for testing, and 12% for validation in Test 3, and its experimental result is indicated in Table 4 with various evaluation metrics. Among all the tests Support Vector Machine and Polynomial regression models were found to be better and hence can be assumed to give good results for long-term prediction in terms of producing small errors such as in Mean Absolute Error and Mean Squared Error in a different set of tests.

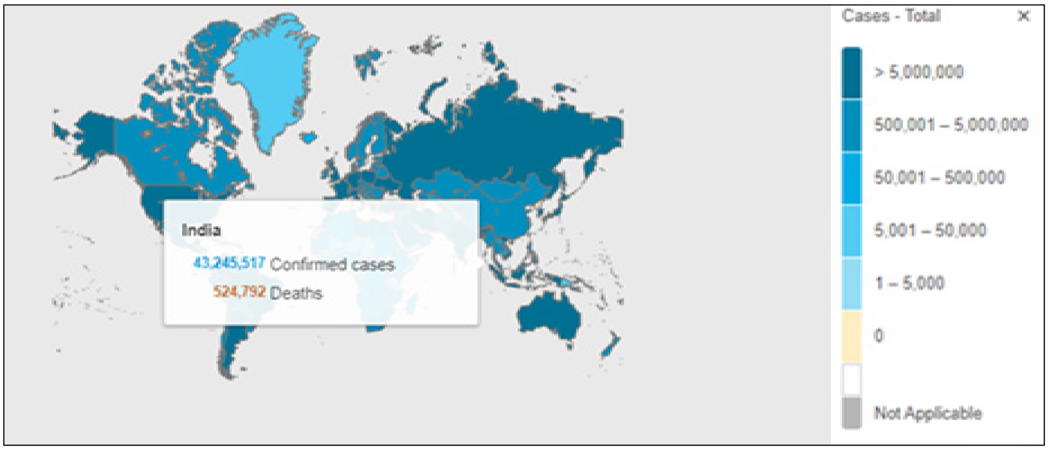


Figure 14. Global visualization

Table 2
Comparison of performance metrics with various models for Test 1

Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R^2 Score	Adjusted R^2 score
Support Vector Regression (Woolf et al., 2021)	3.430898e+07	1.419470e+15	3.767585e+07	0.647203	0.64718
Polynomial Regression (Woolf et al., 2021)	4.988888e+07	4.703992e+15	6.858565e+07	0.312027	0.311991
Bayesian Ridge Polynomial Reduction (Muhammad et al., 2021)	7.051741e+07	5.038594e+15	7.098306e+07	-4.075667	-4.075931
Prophet Prediction	1.081270e+08	1.175406e+16	1.084161e+08	-10.750889	-10.751499
ARIMA	6.914564e+07	5.011705e+15	7.079340e+07	13.341083	-13.341828

Table 3
Comparison of performance metrics with various models for Test 2

Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R^2 Score	Adjusted R^2 score
Support Vector Regression (Woolf et al., 2021)	3.772292e+07	1.774247e+15	4.212181e+07	0.648596	0.648578
Polynomial Regression (Woolf et al., 2021)	6.025114e+07	6.806394e+15	8.250088e+07	0.201452	0.201411
Bayesian Ridge Polynomial Reduction (Muhammad et al., 2021)	6.893719e+07	4.836871e+15	6.954762e+07	-3.054984	-3.055195
Prophet Prediction	1.068353e+08	1.148542e+16	1.071701e+08	-10.629112	-10.629716
ARIMA	7.480714e+07	5.781636e+15	7.603707e+07	-10.997513	-10.998137

Table 4
Comparison of performance metrics with various models for Test 3

Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R^2 Score	Adjusted R^2 score
Support Vector Regression (Woolf et al., 2021)	5.410193e+07	5.258222e+15	7.251360e+07	0.148236	0.148192
Polynomial Regression (Woolf et al., 2021)	8.040728e+07	1.071893e+16	1.035323e+08	-0.355910	-0.355981
Bayesian Ridge Polynomial Reduction (Muhammad et al., 2021)	4.119488e+07	1.903292e+15	4.362673e+07	-0.399827	-0.399900
Prophet Prediction	9.205843e+07	8.527751e+15	9.234582e+07	-10.125871	-10.126450
ARIMA	6.239042e+07	3.971764e+15	6.302193e+07	-9.794867	-9.795428

CONCLUSION

The proposed system compared with Support Vector Regression, Polynomial Regression, ARIMA, Poly Prophet, and Bayesian Ridge Polynomial Reduction. Based on the comparison of the R^2 Score, the R^2 score of the Support Vector Machine and Polynomial regression models was found to be better and hence can be assumed to give good results for the long-term prediction. On the other hand, ARIMA, Prophet Models, and Bayesian Ridge Reduction models are good for short-term predictions (Muhammad & Al-Turjman, 2021). Here, the ARIMA modeling concerning the changing seasonality of the data is used to create a Machine Learning model in comparison to support vector regressions. The results also showcased the accuracy of SVM along with ARIMA techniques and also helped to understand the latest trends in the pattern of the disease, thus enabling further preventive measures. We can also improve the accuracy and precision of the predictions through these techniques. Additionally, we can add features to implement the predictions on the choropleth global map, providing a better understanding of the disease. In the future, we can have more precise datasets having details about various strains of the virus, and we can implement a strain-specific prediction model for COVID-19 as well in the future.

ACKNOWLEDGEMENT

The authors thank all the colleagues who helped to complete this project successfully at Vellore Institute of Technology, Vellore, India.

REFERENCES

Aarthi, A. D., & Gnanappazham, L. (2018). Urban growth prediction using neural network coupled agents-based cellular automata model for Sriperumbudur taluk, Tamil Nadu, India. *The Egyptian Journal of Remote Sensing and Space Science, 21*(3), 353-362. https://doi.org/10.1016/j.ejrs.2017.12.004

Abedini, M., Ghasemyan, B., & Rezaei Mogaddam, M. H. (2017). Landslide susceptibility mapping in Bijar city, Kurdistan province, Iran: A comparative study by logistic regression and AHP models. *Environmental Earth Sciences, 76*(8), Article 308. https://doi.org/10.1007/s12665-017-6502-3

Allwood, B. W., Koegelenberg, C. F., Ngah, V. D., Sigwadhi, L. N., Irusen, E. M., Lalla, U., Yalew, A., Tamuzi, J. L., McAllister, M., Zemlin, A. E., Jalavu, T. P., Erasmus, R., Chapanduka, Z. C., Matsha, T. E., Fwemba, I., Zumla, A., & Nyasulu, P. S. (2022). Predicting COVID-19 outcomes from clinical and laboratory parameters in an intensive care facility during the second wave of the pandemic in South Africa. *IJID Regions, 3*, 242-247. https://doi.org/10.1016/j.ijregi.2022.03.024

Arabi, Y. M., Murthy, S., & Webb, S. (2020). COVID-19: A novel coronavirus and a novel challenge for critical care. *Intensive Care Medicine, 46*(5), 833-836. https://doi.org/10.1007/s00134-020-05955-1

Bae, S., Sung, E., & Kwon, O. (2021). Accounting for social media effects to improve the accuracy of infection models: Combatting the COVID-19 pandemic and infodemic. *European Journal of Information Systems, 30*(3), 342-355. https://doi.org/10.1080/0960085x.2021.1890530

Bird, J. J., Barnes, C. M., Premebida, C., Ekárt, A., & Faria, D. R. (2020). Country-level pandemic risk and preparedness classification based on COVID-19 data: A machine learning approach. *PLoS ONE, 15*(10), Article e0241332. https://doi.org/10.1371/journal.pone.0241332

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real-time. *The Lancet Infectious Diseases, 20*(5), 533-534. https://doi.org/10.1016/s1473-3099(20)30120-1

Estenssoro, E., Loudet, C. I., Dubin, A., Kanoore Edul, V. S., Plotnikow, G., Andrian, M., Romero, I., Sagardía, J., Bezzi, M., Mandich, V., Groer, C., Torres, S., Orlandi, C., Rubatto Birri, P. N., Valenti, M. F., Cunto, E., Sáenz, M. G., Tiribelli, N., Aphalo, V., Bettini, L., Rios, F. G., & Reina, R. (2022). Clinical characteristics, respiratory management, and determinants of oxygenation in COVID-19 ards: A prospective cohort study. *Journal of Critical Care, 71*, Article 154021. https://doi.org/10.1016/j.jcrc.2022.154021

Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. *Science of the Total Environment, 739*, Article 140033. https://doi.org/10.1016/j.scitotenv.2020.140033

Garrido, J., Martínez-Rodríguez, D., Rodríguez-Serrano, F., Pérez-Villares, J., Ferreiro-Marzal, A., Jiménez-Quintana, M., & Villanueva, R. (2022). Mathematical model optimized for prediction and health care planning for COVID-19. *Medicina Intensiva (English Edition), 46*(5), 248-258. https://doi.org/10.1016/j.medine.2022.02.020

Grasselli, G., Pesenti, A., & Cecconi, M. (2020). Critical care utilization for the COVID-19 outbreak in Lombardy, Italy. *JAMA, 323*(16), Article 1545. https://doi.org/10.1001/jama.2020.4031

He, X., Zhou, C., Wang, Y., & Yuan, X. (2021). Risk assessment and prediction of COVID-19 based on epidemiological data from spatiotemporal geography. *Frontiers in Environmental Science, 9*, Article 634156. https://doi.org/10.3389/fenvs.2021.634156

Klyushin, D. A. (2020). Nonparametric analysis of tracking data in the context of COVID-19 pandemic. In A. E. Hassanien, N. Dey & S. Elghamrawy (Eds.), *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach* (pp. 35-50). Springer. https://doi.org/10.1007/978-3-030-55258-9_3

Li, J., Li, S., Cai, Y., Liu, Q., Li, X., Zeng, Z., Chu, Y., Zhu, F., & Zeng, F. (2020). *Epidemiological and clinical characteristics of 17 hospitalized patients with 2019 novel coronavirus infections outside Wuhan, China*. MedRxiv. https://doi.org/10.1101/2020.02.11.20022053

Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J., Vespignani, A., & Santillana, M. (2020). Real-time forecasting of the COVID-19 outbreak in Chinese provinces: Machine learning approach using novel digital data and estimates from mechanistic models. *Journal of Medical Internet Research, 22*(8), Article e20285. https://doi.org/10.2196/20285

Liu, Q. Y., Kwong, C. F., Zhang, S., & Li, L. (2018, November 4). *A hybrid fuzzy-MADM based decision-making scheme for QoS aware handover*. [Paper presentation]. IET Doctoral Forum on Biomedical Engineering, Healthcare, Robotics and Artificial Intelligence 2018 (BRAIN 2018), Ningbo, China. https://doi.org/10.1049/cp.2018.1728

Looi, M. (2020). COVID-19: Is a second wave hitting Europe? *BMJ, 371,* Article 4113. https://doi.org/10.1136/bmj.m4113

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., ... & Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *The Lancet, 395*(10224), 565-574. https://doi.org/10.1016/s0140-6736(20)30251-8

Mahdavi, M., Choubdar, H., Zabeh, E., Rieder, M., Safavi-Naeini, S., Jobbagy, Z., Ghorbani, A., Abedini, A., Kiani, A., Khanlarzadeh, V., Lashgari, R., & Kamrani, E. (2021). A machine learning based exploration of COVID-19 mortality risk. *PLoS ONE, 16*(7), Article e0252384. https://doi.org/10.1371/journal.pone.0252384

Mogensen, I., Hallberg, J., Björkander, S., Du, L., Zuo, F., Hammarström, L., Pan-Hammarström, Q., Ekström, S., Georgelis, A., Palmberg, L., Janson, C., Bergström, A., Melén, E., Kull, I., Almqvist, C., Andersson, N., Ballardini, N., Bergström, A., Björkander, S., ... & Schwenk, J. M. (2022). Lung function before and after COVID-19 in young adults: A population-based study. *Journal of Allergy and Clinical Immunology: Global, 1*(2), 37-42. https://doi.org/10.1016/j.jacig.2022.03.001

Mudenda, S., Mukosha, M., Mwila, C., Saleem, Z., Kalungia, A. C., Munkombwe, D., Daka, V., Witika, B. A., Kampamba, M., Chileshe, M., Hikaambo, C., Kasanga, M., Mufwambi, W., Mfune, R. L., Matafwali, S. K., Bwalya, A. G., Banda, D. C., Gupta, A., Phiri, M. N., ... & Kazonga, E. (2021). *Impact of the coronavirus disease (COVID-19) on the mental health and physical activity of pharmacy students at the University of Zambia: A cross-sectional study*. MedRxiv. https://doi.org/10.1101/2021.01.11.21249547

Muhammad, M. A., & Al-Turjman, F. (2021). Application of IoT, AI, and 5G in the fight against the COVID-19 pandemic. In F. Al-Turhman (Ed.), *Artificial Intelligence and Machine Learning for COVID-19* (pp. 213-234). Springer. https://doi.org/10.1007/978-3-030-60188-1_10

Olszewski, R., Pałka, P., & Wendland, A. (2021, December 13-16). *Agent-based modeling as a tool for predicting the spatial-temporal diffusion of the COVID-19 pandemic*. [Paper presentation]. 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore. https://doi.org/10.1109/IEEM50564.2021.9672878

Pan, W., Deng, Q., Li, J., Wang, Z., & Zhu, W. (2021, July 18-22). *STSIR: A spatial temporal pandemic model with mobility data-A COVID-19 study*. [Paper presentation]. 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China. https://doi.org/10.1109/IJCNN52387.2021.9533596

Pullano, G., Pinotti, F., Valdano, E., Boëlle, P., Poletto, C., & Colizza, V. (2020). Novel coronavirus (2019-nCoV) early-stage importation risk to Europe, January 2020. *Eurosurveillance, 25*(4), Article 2000057. https://doi.org/10.2807/1560-7917.es.2020.25.4.2000057

Quah, P., Li, A., & Phua, J. (2020). Mortality rates of patients with COVID-19 in the intensive care unit: A systematic review of the emerging literature. *Critical Care, 24*, Article 285. https://doi.org/10.1186/s13054-020-03006-1

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B., Aslam, W., & Choi, G. S. (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE Access, 8*, 101489-101499. https://doi.org/10.1109/access.2020.2997311

Shaukat, K., Masood, N., Shafaat, A., Jabbar, K., Shabbir, H., & Shabbir, S. (2015). Dengue fever in perspective of clustering algorithms. *Journal of Data Mining in Genomics & Proteomics, 6*(3), Article 1000176. https://doi.org/10.4172/2153-0602.1000176

Shilo, S., Rossman, H., & Segal, E. (2020). Axes of a revolution: Challenges and promises of big data in healthcare. *Nature Medicine, 26*(1), 29-38. https://doi.org/10.1038/s41591-019-0727-5

Woolf, S. H., Chapman, D. A., & Lee, J. H. (2021). COVID-19 as the leading cause of death in the United States. *Jama, 325*(2), 123-124. https://doi.org/10.1001/jama.2020.24865

World Health Organization. (2020). *Dashboard of the Coronavirus Disease (COVID-19) Outbreak Situation*. World Health Organization. https://covid19/who.int/

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M., Dahly, D. L., Damen, J. A., Debray, T. P., de Jong, V. M., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., ... & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal. *BMJ, 369*, Article 1328. https://doi.org/10.1136/bmj.m1328

Zhao, D., & Zhang, H. (2022, March 25-27). *Comparison of the SVR and ARIMA models for prediction of daily imported new cases of COVID-19 in Shanghai, China*. [Paper presentation]. 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), Hangzhou, China. https://doi.org/10.1109/CACML55074.2022.00048