Proposing and Optimizing COVID-19 Predictions: A Comprehensive Ensemble Approach for Time Series Forecasting in India

Aakanshi Gupta1, Tooba Khan1, Nidhi Mishra1, Nishtha Jatana2\*, Shaily Malik2, Vaani Garg2

1Department of CSE, ASET, AUUP Noida, INDIA.

Email: [aakanshi@gmail.com](mailto:aakanshi@gmail.com), [tooba153khan@gmail.com](mailto:tooba153khan@gmail.com), [nmishra1@amity.edu](mailto:nmishra1@amity.edu)

2Department of CSE, Maharaja Surajmal Institute of Technology, Delhi, India,

Email: [nishtha.jatana@gmail.com](mailto:nishtha.jatana@gmail.com), [shaily.singh@msit.in](mailto:shaily.singh@msit.in), vaani.garg@msit.in

**Abstract**

The novel coronavirus (COVID-19) has devastated millions of people and is a major threat to world health. The world economy was severely disrupted, millions of people died, and many suffered from severe psychological illnesses. Therefore, by employing time-series forecasting of COVID-19 instances for projecting effective cases using time series models, the government will be better able to deal with emergencies of this kind. Values of confirmed COVID-19 instances extremely close to the actual values may be predicted using time series forecasting data. Therefore, the primary objective of the research is to identify a model that outperforms the current models in terms of prediction and is more helpful for predicting emergencies during an epidemic or a pandemic. In the latest research, time-series models such as LSTM (Long-Short Term Memory), ARIMA (Auto Regression Integrated Moving Average), AR (Auto regression), and proposed ensemble model of Lasso regression and ridge regression with gradient boost as meta model have been studied for better Covid-19 prediction. After computing performance metrics, the root-mean-squared error, or RMSE, and mean absolute error (MAE) of each model were discovered and normalized for evaluation of the performance of the prediction models. The proposed ensemble model was found to exhibit better performance than the other models in terms of prediction accuracy. Later studies will concentrate on creating novel models capable of projecting time series data in line with the trajectory of impending COVID-19 variations.

**Keywords:** ARIMA (Auto Regression Integrated Moving Average), AR (Auto regression), LSTM (Long-Short Term Memory)

1. **Introduction**

The covid-19 virus was first recognized in December 2019, in the city of Wuhan, Hubei province in China. There is a belief that it has its origin in an animal resource, probably bats, and has been transmitted to human beings through some immediate host. The total number of positive covid-19 cases as recorded by the World Health Organization was 259,474 and the total number of fatalities had risen to 259,474 across the world by May 2020.COVID-19 was recognized as an international concern, thereby declared as a public health emergency on the 30th of January 2020. The first coronavirus case was reported on the 30th of January,2020 in the state of Kerala. The deadly virus that created the whole mess alike in the initial year worsened in 2021 when our country was hit by the Delta variant. India went through three major waves of COVID-19, where the wave which was driven by the Delta variant in the months of April-May 2021 was the worst. The hospitals in India were scrambling for oxygen and doctors were burdened beyond capacity. The tally of COVID-19 cases per day had gone maximum to 16 million in April 2021, and India became the second in the list of worst-struck countries of COVID-19, the first being America. This situation was traumatizing as the family members of patients were feeling helpless, trying their best to arrange beds and oxygen cylinders for their relatives, and doctors being overburdened and flummoxed because of the rising number of patients and an insufficient number of resources to meet the needs of the whole population.  As the second wave became less severe our country was struck by the third wave (the Omicron variant of covid-19).

This most recent variant of Covid-19 although less severe than the delta variant still poses a danger to humanity. People suffering from this virus have gone through different kinds of symptoms, while some were asymptomatic, others witnessed its severity, eventually leading to death. COVID-19 is still evolving, the most recent version of this virus is variant EG.5.1 (Eris) which was witnessed in the state of Maharashtra in May 2023.

To learn the trend of rising number of covid-19 cases, and be prepared for the coming situations beforehand, time-series-forecasting has been used to predict the number of estimated covid-19 cases for a certain period in future. ARIMA model outperformed FACEBOOKPROPHET model by its effective evaluations and time series forecasting of covid-19 cases [1]. Rauf et al. [2] found that ARIMA model has enhanced predictions as compared to AR models. LSTM has also exhibited promising results in prediction of covid-19 cases. Sunjaya et. al [3] found that LSTM exhibited better RMSE and MAPE values than ARIMA model and Kirbas et. al [4] found that LSTM gave better predictions than both NARNN (Non-linear Auto-regression Neural networks) as well as ARIMA model. LSTM model exhibited less error in predictions which are as small as 0.005% while ARIMA on the other hand in case of ARIMA exhibit 0.007% and 0.006% error rate. Prajapati et. al [5] depict that with hybrid model, enhanced performance can be observed than the traditional forecasting model wherein they used the ARIMA-NARNN hybrid model that uses both Auto-Regression integrated moving average and Non-linear Auto regression neural networks, wherein NARNN works on making non-linear predictions while ARIMA works on linear ones.

The objectives of this work are as follows:

To find a suitable model that makes the best possible predictions for time series data.

To make a comparative analysis of different machine learning, deep learning and ensemble models through some statistical parameter.

1. The key contributions of this work are as follows:
2. Comparative analysis of time-series forecasting models of COVID-19 that include the ARIMA model, AR model, LSTM model, and the proposed hybrid model, which is an ensemble model of lasso regression and ridge regression with gradient boosting as the meta-model.
3. As for AR (Auto Regression model), an analysis has been done to detect whether there is any autocorrelation in the COVID-19 datasets used.  ADF test (Augmented Dickey-Fuller test) has also been performed to check the stationarity of the dataset. Even the ACF and PACF graphs were plotted which showed the correlation of the present value with the past values and then after confirming the stationarity of the dataset, the model has been implemented.
4. For LSTM, a graph is plotted to compare actual and predicted values of COVID-19 cases.
5. For ARIMA, since in the study it was determined earlier that the datasets are stationary, the parameters for the ARIMA model have been calculated based on ACF and PACF plots and then the model is implemented.
6. For the proposed model, the number of features has been increased by considering the number of cases a day before, 2 days before, and 3 days before the current date. Then, the lasso regression model and the ridge regression model have been applied to the dataset. Then, the meta model i.e. the gradient boosting algorithm has been applied to the predicted values of these two models.
7. RMSE (Root mean square error) values and mean absolute error values have been calculated for each model and the comparison is done.
8. The rest of the paper is structured as follows.  Section 2 represents the literature review, section 3 represents methodology, section 4 contains the proposed model section 5 illustrates performance metrics section 6 refers to results followed by Section 7 as a discussion section containing threats to validity, and Section 9 consists of conclusion and future work.

**Related Work**

Deep learning algorithms and regression models have been investigated for their performance evaluation on time-series prediction [6]. COVID-19 first originated in Wuhan, a Hubei province at the beginning of December 2019 where several patients suffered from this contagious viral infection which later, on the 30th of January 2020 was declared as a matter of international health concern by the World Health Organization (WHO). It is a highly infectious disease that caused such a horrendous situation that people lost their lives not only because of being afflicted with the virus but also because of the financial crisis they suffered as well as their disturbed mental state. Because of the rising number of cases, India was forced to impose a lockdown from 24th of March to 31st of May 2020. Covid-19 cases reached 1,01,139 on 18th of May in India. In the past, many authors have proposed techniques to optimize the predictions of COVID-19 [22, 23]

ARIMA has been used for time series prediction for a long time.Conejo.et.al[39] used ARIMA for electricity price forecasting while Wadi. et.al[40] used it for predicting financial time series data.

ARIMA model gave promising results in time series forecasting as compared to FACEBOOKPROPHET [1]. Khan et. al [7] used ARIMA (1,1,0) along with NAR (Non-linear autoregressive neural networks), wherein it was found to exhibit the highest R2 value of 0.95. ARIMA has three major components, namely AR (Auto regression), MA (moving average) and I (integration). ARIMA and its constituents were observed to statistically visualize the Covid-19 dataset and analyze its pattern. ARIMA has been hybridized with NARNN (nonlinear autoregression neural network). A hybrid model of ARIMA with neural networks has also been studied for time-series prediction by Zhang et al. [8]. The use of ARIMA models can also be vindicated by its usage in Yamak et. al [9], wherein it was shown to outperform advanced deep learning models like LSTM and GRU.

|  |  |  |  |
| --- | --- | --- | --- |
| **Reference no.** | **Methodology** | **Results** | **Limitations** |
| [11] | Three variants of Recurrent Neural Networks (RNN) such as Stacked LSTM, Bi-directional LSTM and ConvLSTM. | Convolution LSTM outperformed the other two models and predicts the Covid-19 cases with high accuracy and very less error for all four datasets of both countries | To forecast possible Covid-19 cases for other countries and also aerosol transmission of Covid-19 can be verified. |
| [20] | This paper presents a comparative study of five deep learning methods to forecast the number of new cases and recovered cases. RNN, LSTM, BiLSTM, GRUs and (VAE) algorithms have been applied for global forecasting. | Their Results have successfully demonstrated the promising potential of the deep learning model in forecasting COVID-19 cases and highlight the superior performance of the VAE compared to the other algorithms. | All other models except VAE perform moderate forecasting performance. This is maybe due to their need for more data in the training to capture the dynamics of COVID-19. |
| [1] | They have adopted ARIMA and Facebook Prophet (FBProphet) model in their evaluative and forecasting study. | The results show that ARIMA model is more effective for forecasting COVID-19 prevalence. The forecasting results have potential to assist governments to plan policies to contain the spread of the virus. | Further can be improved by taking various variables into account like population density, weather, health system, patient history etc. using DL and AI. |
| [7] | ARIMA and NAR models have implemented. | The results showed an increasing trend in the actual and forecasted numbers of COVID-19 cases with approximately 1500 cases per day, based on available data as on 04th April 2020. | However, predictions can be improved by taking a few preventive steps; by collecting more data from the upcoming days. For future modifications, the new methodology will be used. |
| [21] | ARIMA, AR and MA models have implemented. | This paper has successfully concluded Time Series Forecasting model and has analysed the COVID-19 epidemic occurrence. | Testing rates are less due to being in first stage of COVID-19 outbreak in India. |
| [2] | LSTM, GRU and RNN | They have utilized latest deep learning algorithms to quantify the intensity of pandemic for near future by employing time variable and data non-linearity while employing neural networks. | Results can be expanded by monitoring the spread of the pandemic in the Gulf and some European countries and can be improved by extensive study in our future research. |
| [3] | ARIMA and LSTM approaches | LSTM gave better results than ARIMA, BASED ON RMSE and MAPE values. | Not able to forecast covid-19 cases between 2nd March 2020 till 31st March 2020. |
| [4] | Auto-Regressive Integrated Moving Average (ARIMA), Nonlinear Autoregression Neural Network (NARNN) and Long-Short Term Memory (LSTM) approaches. | Among all 6 methodologies LSTM was found to show maximum accuracy. | Restricted amount of covid-19 is present which is troublesome for their modelling. |
| [10] | LSTM, ARIMA | Errors in predictions were calculated using parameters like KMAPE and KMDSA for LSTM models which came as less as 0.05 percent while in case of ARIMA these parameters came 0.07 and 0.06. | Other time series forecasting models can also be evaluated and developed as we collect more data. |
| [5] | ARIMA, LSTM, Holt Winters, Prophet, ARIMA-NARNN model | ARIMA-NARNN hybrid model performed better than all the other models. | Other time series forecasting models can also be evaluated and developed as we collect more data. |
| [41] | Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells, autoregressive Integrated moving average (ARIMA), seasonal autoregressive Integrated moving average (SARIMA) for forecasting COVID-19 trends | Here for most of the countries LSTM and GRU excelled regression based models, namely ARIMA and SARIMA. | Here the analysis is country specific but top 10 countries, in terms of highest number of covid-19 cases were considered and only two metrics were considered for evaluation namely MSE and RMSE. |
| [42] | Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning | Covid-19 Cases of three months were predicted i.e from 7th April 2020 to 7th July 2020 using ARIMA ,which gave high accuracy with a confidence level of 95% | These predictions will only be valid if the nature of this covid-19 pandemic remains same. |
| [43] | Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models | Here ARIMA model trained on observed covid-19 data from 22 January 2020 till 31st March 2020 was tested against observed cases from 18th April 2020 to 1st May 2020 and it gave a MAPE value of 16.01. | Predictions were only restricted to Malaysia. |
| [44] | Progression of COVID-19 Cases in  Telangana State by using ARIMA, MLP,  ELM and LSTM Prediction Models by  Retrospective Confirmation | LSTM gave excellent results in comparison to ELM,MLP and ARIMA with LSTM giving least value of RMSE as compared to other models. | Predictions were only made for Telangana. |
| [45] | Predicting the Spread of a Pandemic Using Machine Learning: A Case Study of COVID-19 in the UAE | Here univariate LSTM gave the best results as compared to Lasso regression and Exponential Smoothing. | Since data is purely numerical and statistical in nature therefor, in depth research could’nt be made and the vaccination count was also not considered. |
| [46] | LSTM algorithm optimization for COVID-19 prediction model | Pop LSTM gave better results as compared to Improved Minmax Scaler LSTM and basic LSTM | The difference between data and data spacing which makes it necessary to make adjustments to the model during the prediction process. |
| [47] | COVID-19 prediction using LSTM algorithm: GCC case study | Here LSTM showed that Qatar and KSA will take larger time to recover from Covid-19. | Factors like age, social distancing and  Weather were not considered while making predictions. |

**Table I: Review of related work**

LSTM is a type of recurrent neural network that can handle “long-term dependencies”, and is used by many researchers to perform time-series forecasting of Covid-19 cases. LSTM generally remembers the past values and, in each cell state, the content to be remembered is filtered and the information to be added is decided. This concept even overcomes the problem of vanishing and exploding gradient faced by RNN (Recurrent neural networks). LSTM being a deep learning model generally outperforms regressive models in time series forecasting as shown by Barman et al. [10] in which he compared LSTM with ARIMA models. LSTM outperforms other deep-learning neural networks like GRU and RNN [2]. It was observed that Convolutional LSTM among all types of LSTM gave the best results for time-series forecasting of COVID-19 cases [11-13] and even outperforms the ARIMA Model [14]. LSTM has also shown promising results in other fields of forecasting [9][15].

Substantial research has been conducted in time series forecasting of COVID-19 cases, where every study used unique methods for forecasting and versatile parameters for forecasting. Some of the related work is mentioned in Table I.

Many ensemble models have been created for time series forecasting in the past. Corizzo. et.al[48] used it for stock market prediction while Iftikhar. et.al[49] used it for electricity demand forecasting. But none have been made for Covid-19 prediction.

1. **Methodology**

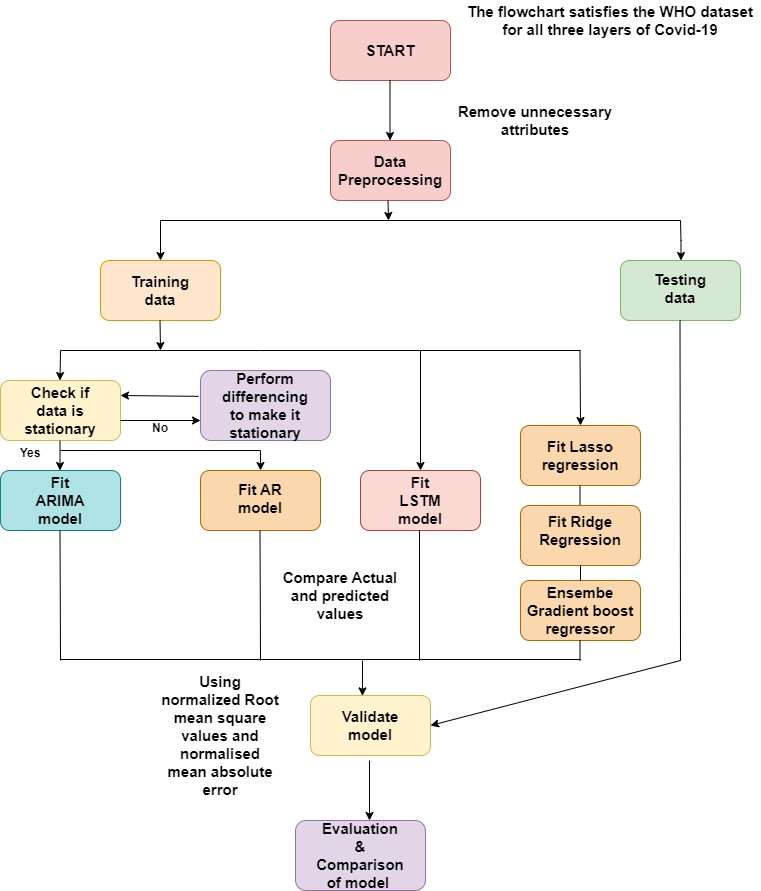
Time series data is recorded using several observations recorded over a continuous period. As for time-series forecasting, it is a practice to predict values over a period in the future [16]. Three major components are of interest when doing time-series forecasting.

1.)    Seasonality: Period when the observations captured are at their peak. These are certain periods where some irregularities are differentiating from the normal trend.

2.)    Trend: Gives us if there is an increasing or decreasing pattern or trend in observed values.

3.)    Unexpected events (also known as noise or irregularities): Some dynamic events that occur unexpectedly.

This work deals with time-series forecasting, where the time-series data was pre-processed, and thereby, various models were evaluated for its prediction accuracy.



**Fig 1:** Flow-chart of the proposed work

Fig. 1 Sequence of the conducted work. It has been kept in mind that all the necessary operations are performed for the conversion of data into the required form if not present and to conduct required tests and plot required graphs to support the model being implemented.

* 1. **Dataset Collection**

Covid-19 data repository by WHO (https://data.who.int/dashboards/covid19/data?n=c) which has its time series from 11th March 2020 to 28th February 2022. The dataset has then been grouped based on time series into categories namely, the first wave of COVID-19, the second wave (delta variant), and the third wave(omicron) in India. Heatmaps are a way of representing data graphically such that the values are represented as cells individually and are presented through various colors. In the WHO heatmap, it can be observed that as the color changes from black to light yellow, there is a rise in the number of cases as shown in Fig. 2.A chart with numbers and a number of cases

Description automatically generated with medium confidence

**Fig 2:** Correlation heatmap for WHO dataset

Fig. 2 shows the correlation heat map which is a 2-dimensional matrix such that there are two discrete dimensions for representation of data [17]. Here we use colored cells to present each cell such that the color is proportional to the number of measurements that match the dimension values. The above two figures imply the pattern and variation in the Covid-19 data especially the number of confirmed Covid-19 cases which is the need of the hour.

1. **Proposed Model**

This section explains the proposed model which is an Ensemble model of Lasso Regression, Ridge Regression, and Gradient boosting as a meta-model that is used for Stacking.

The ensemble model has been formed by stacking of lasso regression model and ridge regression model, using the Gradient boosting algorithm as the meta-model for the ensemble model. The baseline models are explained as follows:

1. **Lasso regression:** which has its full form as “Least Absolute Shrinkage and Selection Operator”, which is known for performing L1 regression. This leads to a smaller slope line by adding a penalty. Here the keywords that can be derived from its full form are absolute and selection.

***Target=******y-axis-intercept + slope \* response\_variable [I]***

***Which minimizes ,***

***The sum of squared residuals + λ \*|slope|***

***Here λ can be from 0 to +ve infinity***

1. **Ridge regression:** This model just like lasso regression helps to prevent overfitting by adding a penalty but here this regression performs L2 regularization. Here instead of taking absolute of coefficients as the penalty, the squares of those coefficients are taken. Thus, it makes the target variable less sensitive to response variables by fitting a smaller slope line to the datasets.

***Target=y-axis-intercept + slope\*response\_variable [II]***

***it minimizes***

***The sum of squared residuals+ λ \*slope^2,***

***Here λ can be from 0 to +ve infinity***

**3) Gradient descent Boosting algorithm:** This technique used as a metamodel, usually ensembles the predictions made by various weak learners in sequence and thus aims to provide better predictions. This is done by building models sequentially and thus each model tends to improve the previous model’s predictions. Thus, when the target variable has continuous values, this algorithm is known as a Gradient-boosting regressor. Thus, the main aim of this regressor algorithm is to minimize the loss functions by gradient descent algorithm.

The algorithm of the proposed model is as follows:

* Important libraries imported
* CSV file is read
* Increasing the number of features by adding the number of cases a day before,2 days before, and 3 days before in the dataset.
* Splitting the dataset into training and testing sets where the last 30 days' dataset has been made the testing dataset.
* Defining the baseline models and then ensembling them through gradient boosting regressor as follows:

5.1) Function lasso regression(x,y)

**y=y-axis-intercept + slope \* x [III]**

**which minimizes,**

**The sum of squared residuals + λ \*|slope|,**

**Here λ can be from 0 to +ve infinity**

5.2) Function Ridge regression(x,y)

**y=y-axis-intercept + slope \* x [IV]**

**which minimizes,**

**The sum of squared residuals + λ \*slope2**

**Here λ can be from 0 to +ve infinity**

1. The predictions made on training dataset has been have then been made to train the meta model which is gradient ensemble model.

Function gradient descent boost regressor()

**yp,i= slope\*x + ya,I [V]**

**Now minimising the residual distance by rmse,**

**[VI]**

**Calculating pseudo residuals by considering tree[y\_predicted-y\_actual]**

**[VII]**

Here,T(xi) is the predictions made by previous tree.

1. Calculating the metrics:

**MAE=mean\_absolute\_error(Y\_new\_pred,y\_test) [VIII]**

**RMSE=mean\_squred\_error(Y\_new\_pred,y\_test) [IX]**

Table II indicates that the proposed ensemble model has never been implemented and that none of the above research studied the three layers of covid-19 individually.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Title** | **WHO**  **dataset** | **ARIMA Model Implementation** | **AR Model** | **LSTM Model** | **Ensemble model of Lasso regression and ridge regression** | **Comparative analysis** |
| Proposed model | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India [7] | 🗴 | ✓ | 🗴 | 🗴 | 🗴 | 🗴 |
| Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches [4] | 🗴 | ✓ | 🗴 | ✓ | 🗴 | 🗴 |
| Forecasting of Covid-19 positive cases in Indonesia using long short-term memory (LSTM)  (Sunjaya [3] | 🗴 | ✓ | 🗴 | ✓ | 🗴 | 🗴 |
| Time Series Analysis and Forecasting of COVID-19 Cases  Using LSTM and ARIMA Models  [10] | 🗴 | ✓ | 🗴 | ✓ | 🗴 | 🗴 |

**Table II:** Comparative evaluation of the existing literature and the proposed ensemble approach

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| COVID-19 Pandemic Prediction using Time Series Forecasting Models [1] | 🗴 | ✓ | 🗴 | 🗴 | 🗴 | 🗴 |
| Forecasting COVID-19 Pandemic Using Prophet, ARIMA, and Hybrid Stacked LSTM-GRU Models in India [18] | 🗴 | ✓ | 🗴 | ✓ | 🗴 | 🗴 |
| Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods [19] | 🗴 | ✓ | 🗴 | 🗴 | 🗴 | 🗴 |
| Comparison of Traditional and Hybrid Time Series  Models for Forecasting COVID-19 Cases [5] | 🗴 | ✓ | 🗴 | 🗴 | 🗴 | 🗴 |
| Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study [20] | 🗴 | 🗴 | 🗴 | ✓ | 🗴 | 🗴 |
| Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks [2] | 🗴 | 🗴 | 🗴 | ✓ | 🗴 | 🗴 |
| COVID-19 Outbreak: An Epidemic Analysis using Time Series Prediction Model  [21] | 🗴 | ✓ | ✓ | 🗴 | 🗴 | 🗴 |
| DL methods for forecasting COVID-19 time-Series data:A Comparative study [18] | 🗴 | 🗴 | 🗴 | ✓ | 🗴 | 🗴 |
| Time series forecasting of COVID-19 transmissionAsia Pacific countries using DNN  [2] | 🗴 | 🗴 | 🗴 | ✓ | 🗴 | 🗴 |
| COVID-19: An Epidemic Analysis using Time Series Prediction Model [21] | 🗴 | ✓ | ✓ | 🗴 | 🗴 | 🗴 |

1. **Experimental Results**

Comparative evaluation of the time series prediction of COVID-19 of the existing techniques with the proposed models has been made to determine the accuracy of these models. The models used for evaluation include ARIMA (Auto-regression integrated moving average model), AR (Auto-regression model), LSTM (Long-short term memory model), and the ensembled model of lasso and ridge regression.

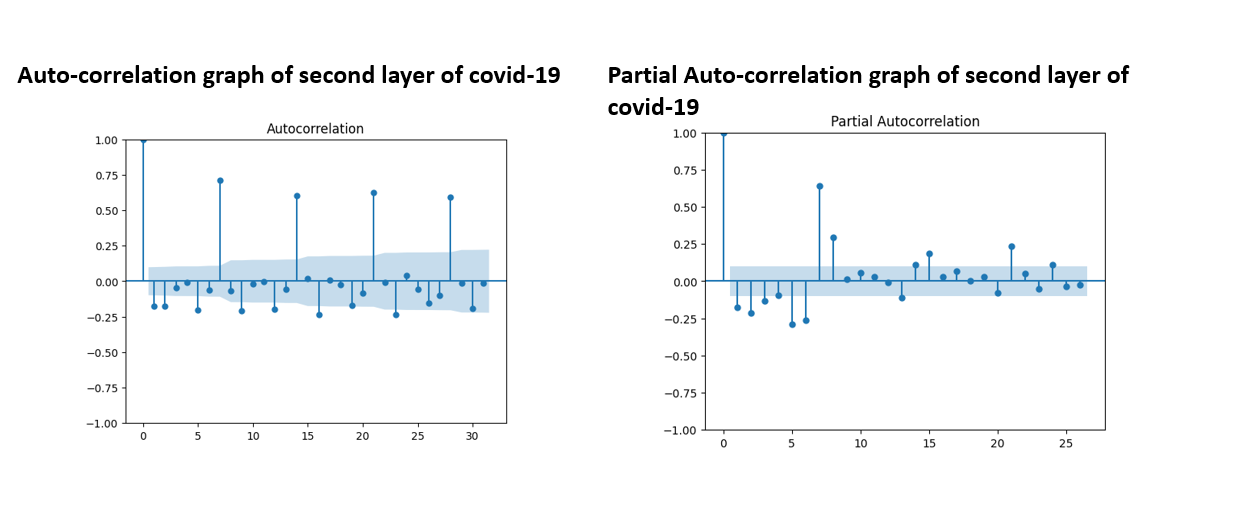
As for ARIMA and AR models, to check if the data is stationary or not, the ADF test was performed. If the p-value comes to be less than 5% then the data is stationary. To further confirm the stationarity of the dataset, ACF (Autocorrelation plot) and PCF (Partial Autocorrelation plot) were used. Before discussing PCF, it is essential to know the correlation between two variables. Thus, PCF explains how much two values are correlated after the intermediate lag values are discarded.

The results of the ADF test for the datasets of three waves of COVID-19 in India have been different. If any of the dataset values were not stationary, the dataset was then differenced until it became stationary. The final values of the ADF test for the dataset of all three layers of COVID-19 were found to be less than 5%. Then the model was implemented, and thereafter, the confirmed cases of COVID-19 were predicted.

To fit the ARIMA model, it is essential to find the number of lags(p), the number of times differencing needs to be performed(d), and the order of moving average(q). Thus, after observing the ACF and PACF plots, the value of p was actually decided from the PACF plot, by observing the lag(n) which is very close to the dense part, and then initializing the value of p by (n-1), as for the value of d it is the number of times the dataset has been differenced to make it stationary and as for the value of q it is decided from the ACF plot by again observing the lag closest to the dense part and then initializing it with (n-1).A graph of covid-19 and partial auto-correlation graph

Description automatically generated

**Fig. 3.** ACF and PCF plot for first wave of covid-19



**Fig. 4.** ACF and PCF plot for second wave of covid-19

A comparison of a graph

Description automatically generated

**Fig.5.** ACF and PCF plots of third wave of covid-19

Here Fig. 3, Fig. 4, and Fig. 5 imply that the datasets for Covid-19 cases which had seasonality such that their variance and mean kept on changing, were differenced to make them stationary. Then on plotting their ACF PACF plot the stationarity of the dataset is verified. Since there is a significant value of correlation at each lag and there is a geometrical decrease in correlation, the partial auto-correlation plot implies that the data after being pre-processed became stationary to be used with AR and ARIMA models. After all this preprocessing of data, the ARIMA model was then been implemented on the stationary dataset and the required predictions were made, as shown in Fig. 6.

1. **Discussion**

Considering the results of the applied models, the proposed ensemble model which is the combination of lasso regression and ridge regression gave the best results as it gave the smallest values of RMSE and MAE in comparison to the other models. But it can be said, that all the models gave comparable results, such that the main purpose of predicting such values which are very much near to the actual cases of COVID-19 is achieved.

As for the AR model, before fitting the model, the value of p (the number of lags to be considered) needs to be decided after observing the ACF plot, which here was tried up to a combination of 15 lags and the order (number of lags) of seasonality was taken to be 13. Thus, here the auto regression model was able to adequately satisfy the trend and seasonality of the dataset and make accurate predictions.

As for the ARIMA model, after finding suitable values of p, d, and q which came out to be (1,1,0) for all three layers of the WHO dataset, which was applied only after making these datasets stationary, proximity was seen between the values it predicted and the actual values.

Now, for LSTM, the required number of epochs the satisfiable value of several features, and the number of hidden layers were decided before the implementation such that the model neither underfits nor overfits the data and accurate predictions can be made. Because of its trait of capturing “long-term dependencies”, it filters the past data and takes into account the effective data, discarding the rest in each cell state and thus making remarkable predictions.

As for the ensemble model, the lasso and ridge regressions which are known to give better results than all other regression models when combined, and backed up by gradient boosting gave the best possible results of all the other methods in this study.

1. **Threats to validity**

The section presents the threats to the validity of the proposed work.

1. History:  It is a very crucial threat to the validity of proposed research imposed due to any chance of an event that affects the observed time series. Since the population per square feet area, the severity of other diseases people were afflicted with, the standard of living of different societies of people, and their hygiene are not considered in the study which may have impacted the time-series data, the history can be misconstrued.
2. Evaluation metrics used: This refers to another important threat that tells us the changes as to how the time series is measured and how the module is evaluated. Here there can be two scenarios where this type of threat can occur. First is a smaller number of evaluation metrics of models and second is a smaller number of researches using normalized Root-mean square error as a parameter for comparison of models.
3. Testing: This type of threat can generally occur when the outcome of an event along with the treatment being applied concurrently is measured. Here the use of vaccine, the effect of lockdown, the necessary hygiene steps, and the social distance practices were not considered while predicting the values.
4. Data: The limited amount of COVID-19 data due to less number of testing in India is a major problem.

While these issues hold some importance, but discussion of these is beyond the scope of this study.

1. **Conclusion and Future Work**

Covid-19 has created havoc in the lives of millions of people across the globe. With the advent of vaccines, the situation has been much better, but the virus is still prevalent and is still mutating. Thus COVID-19 predictions based on time-series forecasting may help in a larger way to predict and thereby prevent emergencies.

The proposed work is supported by statistical tests and visualizations to support the results of our proposed ensemble model. All the predictions, made so far by different proposed methodologies in this research are quite accurate. The parameters like several epochs and batch size input features were decided carefully for each model so that they make the required predictions effectively.

In the future, the proposed model can be further modified for improved accuracy. Newer and more efficient hybrid models combining one or more machine learning models can be evaluated. In the future, newer models will be able to perform time series forecasting according to the pattern of the future COVID-19 variants, which generally come in a unique form having different traits. Also, future work must focus on using more lightweight models with better predictions and on models that are capable of automatic machine adjustments which can help to reduce training time.

In the future researchers can consider vaccination as another factor considering total vaccination count and further consider the type of vaccination dose whether the person, after having vaccination was afflicted with COVID-19 or not. Age, weather, additional lockdowns, and social distancing are other factors that researchers can consider in the future. Further, we can make predictions about specific cities also which can indicate the extent of the spread of covid-19 in an area.

In the future, better research can be made with a dataset considering more factors leading to the spread of this pandemic.

In addition, a strategy to manage uncertainty must be designed to quantify uncertainty and provide more genuine information to users.

**9. Compliance with Ethical Standards:**

**Disclosure of potential conflicts of interest:** No Conflict of interest to disclose.

**Research involving Human Participants and/or Animals:** None

**Informed consent:** Not Applicable

**References**

1. Bontempi, G., Taieb, S., & Borgne, Y. Le. (2013). Machine learning strategies for time series forecasting. *Business Intelligence*, 62–77. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-36318-4\_3
2. Mukhopadhyay, U.; Skjellum, A.; Hambolu, O.; Oakley, J.; Yu, L.; Brooks, R. A brief survey of Cryptocurrency systems. In Proceedings of the 14th Annual Conference on Privacy, Security and Trust (PST), Auckland, New Zealand, 12–14 December 2016; pp. 745–752. [CrossRef]
3. Mahir Iqbal1, Muhammad Shuaib Iqbal2, Fawwad Hassan Jaskani1,\*, Khurum Iqbal2 and Ali Hassan2 Time-Series Prediction of Cryptocurrency Market using Machine Learning Techniques doi: 10.4108/eai.7-7-2021.170286
4. Josh Hawarth; How Many Cryptocurrencies are there <https://explodingtopics.com/blog/number-of-cryptocurrencies>
5. Statista: Crypto currency market Caps <https://www.statista.com/chart/23160/biggest-crypto-currency-market-caps/>
6. A Novel Cryptocurrency Price Prediction Model Using GRU,LSTM and bi-LSTM Machine Learning Algorithms Mohammad J. Hamayel and Amani Yousef Owda
7. Time series analysis of Cryptocurrency returns and volatilities ; Rama K. Maladi, Prakash Dhairiya
8. Ioannis E. Livieris 1,\* , Emmanuel Pintelas 1, Stavros Stavroyiannis 2 and Panagiotis Pintelas 1: Ensemble Deep Learning Models for Forecasting Cryptocurrency Time-Series
9. Chen Peng, Guo Yichao : Cryptocurrency Price Analysis and Time Series Forecasting
10. Time Series Classification of Cryptocurrency Price Trend Based on a Recurrent LSTM Neural Network : Do-Hyung Kwon\*, Ju-Bong Kim\*\*, Ju-Sung Heo\*, Chan-Myung Kim\*\*\*, and Youn-Hee Han\*\*
11. Kaggle: <https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data>
12. Shradha Goled: Why are people Bashing Facebook Prophet : <https://analyticsindiamag.com/why-are-people-bashing-facebook-prophet/>
13. Yugesh Verma: Complete Guide to Bi-Directional LSTM: <https://analyticsindiamag.com/complete-guide-to-bidirectional-lstm-with-python-codes/>
14. Elton: Gated Recurrent Units (GRUs) for Natural Language Processing: <https://pythonwife.com/gated-recurrent-units-grus-for-natural-language-processing/>
15. <https://www.researchgate.net/figure/VAR1-Schematic-of-relationships-modeled-by-vector-autoregression-at-a-lag-of-one_fig1_343442292>
16. or Azizah Hitam, Amelia Ritahani Ismail; Comparative Performance of Machine Learning Algorithms for Cryptocurrency Forecasting
17. Xiao Chan ; P2P Crypto Exchange Script - Create a P2P Cryptocurrency Exchange; <https://www.finextra.com/blogposting/21746/p2p-crypto-exchange-script---create-a-p2p-cryptocurrency-exchange>
18. Jake Frankenfield ; Peer-to-Peer, 2021; <https://www.investopedia.com/terms/p/ptop.asp>
19. Bitcoin documentation: <https://bitcoin.org/en/>

# Guorui Li,  Ying Wang : Automatic ARIMA modeling-based data aggregation scheme in wireless sensor networks

1. Neha Bora: Understanding ARIMA Models for Machine Learning: <https://www.capitalone.com/tech/machine-learning/understanding-arima-models/>
2. CFI team: Autoregressive Integrated Moving Average (ARIMA): <https://corporatefinanceinstitute.com/resources/knowledge/other/autoregressive-integrated-moving-average-arima/>
3. Ajay Ohri: Understanding GRU in 2021: <https://www.jigsawacademy.com/blogs/data-science/gru/>
4. Xin Wang *et al* 2019 *J. Phys.: Conf. Ser.* 1325 012089: OGRU an Optimised Gated Neural Network: <https://iopscience.iop.org/article/10.1088/1742-6596/1325/1/012089/pdf>
5. Gers, F.A., Eck, D., Schmidhuber, J. (2002). Applying LSTM to Time Series Predictable Through Time-Window Approaches. In: Tagliaferri, R., Marinaro, M. (eds) Neural Nets WIRN Vietri-01. Perspectives in Neural Computing. Springer, London. https://doi.org/10.1007/978-1-4471-0219-9\_20
6. Pritika Bahad, Preeti Saxena, Raj Kamal,Fake News Detection using Bi-directional LSTM-Recurrent Neural Network, Procedia Computer Science, Volume 165, 2019, Pages 74-82, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2020.01.072>.
7. (2006). Vector Autoregressive Models for Multivariate Time Series. In: Modeling Financial Time Series with S-PLUS®. Springer, New York, NY. <https://doi.org/10.1007/978-0-387-32348-0_11>
8. N. Kumar and S. Susan, "COVID-19 Pandemic Prediction using Time Series Forecasting Models," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225319.
9. S.L. Ho, M. Xie, The use of ARIMA models for reliability forecasting and analysis, Computers & Industrial Engineering, Volume 35, Issues 1–2, 1998, Pages 213-216, ISSN 0360-8352, <https://doi.org/10.1016/S0360-8352(98)00066-7>
10. Liu, B.; Fu, C.; Bielefield, A.; Liu, Y.Q. Forecasting of Chinese Primary Energy Consumption in 2021 with GRU Artificial Neural Network. *Energies* 2017, *10*, 1453. <https://doi.org/10.3390/en10101453>
11. Huaizhi Wang, Zhenxing Lei, Xian Zhang, Bin Zhou, Jianchun Peng, A review of deep learning for renewable energy forecasting, Energy Conversion and Management, Volume 198, 2019, 111799, ISSN 0196-8904, <https://doi.org/10.1016/j.enconman.2019.111799>
12. S. Siami-Namini, N. Tavakoli and A. Siami Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 1394-1401, doi: 10.1109/ICMLA.2018.00227.
13. Farah Shahid, Aneela Zameer, Muhammad Muneeb, Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM, Chaos, Solitons & Fractals, Volume 140, 2020, 110212, ISSN 0960-0779, <https://doi.org/10.1016/j.chaos.2020.110212>
14. Christopher A. Sims, 1986. "[Are forecasting models usable for policy analysis?](https://ideas.repec.org/a/fip/fedmqr/y1986iwinp2-16nv.10no.1.html)," [Quarterly Review](https://ideas.repec.org/s/fip/fedmqr.html), Federal Reserve Bank of Minneapolis, vol. 10(Win), pages 2-16.
15. Oyedele, Azeez A., et al. “Performance Evaluation of Deep Learning and Boosted Trees for Cryptocurrency Closing Price Prediction.” *Expert Systems with Applications*, vol. 213, Mar. 2023, p. 119233, doi:10.1016/j.eswa.2022.119233.
16. Seabe, Phumudzo Lloyd, et al. “Forecasting Cryptocurrency Prices Using LSTM, Gru, and Bi-Directional LSTM: A Deep Learning Approach.” *Fractal and Fractional*, vol. 7, no. 2, 18 Feb. 2023, p. 203, doi:10.3390/fractalfract7020203.
17. Toai, Tran Kim, et al. “Arima for Short-Term and LSTM for Long-Term in Daily Bitcoin Price Prediction.” *Artificial Intelligence and Soft Computing*, 2023, pp. 131–143, doi:10.1007/978-3-031-23492-7\_12.
18. Zhong, Chao, et al. “LSTM-Regat: A Network-Centric Approach for Cryptocurrency Price Trend Prediction.” *Decision Support Systems*, vol. 169, June 2023, p. 113955, doi:10.1016/j.dss.2023.113955.

[39] Conejo, Antonio J., et al. "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models." *IEEE transactions on power systems* 20.2 (2005): 1035-1042.

[40] Al Wadia, M. T. I. S., and M. Tahir Ismail. "Selecting wavelet transforms model in forecasting financial time series data based on ARIMA model." *Applied Mathematical Sciences* 5.7 (2011): 315-326.

[41] ArunKumar, K. E., et al. "Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells, autoregressive Integrated moving average (ARIMA), seasonal autoregressive Integrated moving average (SARIMA) for forecasting COVID-19 trends." *Alexandria engineering journal* 61.10 (2022): 7585-7603.

[42] Chyon, Fuad Ahmed, et al. "Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning." *Journal of virological methods* 301 (2022): 114433.

[43] Chyon, Fuad Ahmed, et al. "Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning." *Journal of virological methods* 301 (2022): 114433.

[44] Singh, Sarbhan, et al. "Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models." *Journal of infection in developing countries* 14.9 (2020): 971-976.

[45] Sankalpa, Donthi, et al. "Predicting the Spread of a Pandemic Using Machine Learning: A Case Study of COVID-19 in the UAE." *Applied Sciences* 14.10 (2024): 4022.

[46] Sembiring, Irwan, Sri Ngudi Wahyuni, and Eko Sediyono. "LSTM algorithm optimization for COVID-19 prediction model." *Heliyon* 10.4 (2024).

[47] Ghany, Kareem Kamal A., Hossam M. Zawbaa, and Heba M. Sabri. "COVID-19 prediction using LSTM algorithm: GCC case study." *Informatics in Medicine Unlocked* 23 (2021): 100566.

[48] Corizzo, Roberto, and Jacob Rosen. "Stock market prediction with time series data and news headlines: a stacking ensemble approach." Journal of Intelligent Information Systems 62.1 (2024): 27-56.

[49] Iftikhar, Hasnain, et al. "Electricity Demand Forecasting Using a Novel Time Series Ensemble Technique." IEEE Access (2024).