International Journal of
Electrical and Computer System Design
ISSN: 2582-8134
www.ijecsd.com

# Covid-19 Data Analysis And Forecasting

Apirajitha P S*[1], Mahalakshmi G[2], Durai Murugan M[3],

[*1] Department of CSE, Sri Venkateswara college of engineering, Sriperumbudur, chennai, India, apirajithaps@svce.ac.in.

[2] Department of Information Science & Technology, College of Engineering, Anna University, Guindy, Chennai, mlakshmig27@gmail.com.

[3] Department of Information Science & Technology, College of Engineering, Anna University, Guindy, Chennai, duraimurugan@auist.net.

***Abstract:*** *A basic premise of machine learning is to create an algorithm that can take input data, update the output when new data is available, and use statistical analysis to predict the output. Long-term memory is a type of recurrent neural network. Forecasting is an additive, model-based, time-series data forecasting process that adjusts for non-linear trends for year, week, and day seasonality and holiday effects. This works best for time series with strong seasonal influences and for multi-year historical data. Strong prediction of missing data and trend changes, usually with graceful exception handling. Autoregressive Integrated Moving Average (ARIMA) is a statistical analysis model that uses time series data to better understand a data set and predict future trends. Statistical models are autoregressive when they predict future values based on past values.*

**Keywords:** Machine Learning, LSTM, Prophet.

## 1. Introduction

Without being explicitly coded, software systems can predict events more accurately thanks to a class of algorithms known as machine learning (ML). Building algorithms that can take input data and apply statistical analysis to predict an output while updating outputs as new data becomes available is the fundamental idea behind machine learning. Recurrent neural networks are a type of long short term memory. The output from the previous phase is sent into the current step of an RNN as input. It addressed the issue of long-term RNN dependency, in which the RNN can predict words from current data but cannot predict words held in long-term memory. RNN's performance becomes less effective as the gap length increases. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly,and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. When handling outliers, Prophet often does a good job of handling missing data and changes in the trend. In order to comprehend a data set more fully or to anticipate future patterns, statistical analysis models called autoregressive integrated moving averages, or ARIMAs, use time series data. If a statistical model forecasts future values using data from the past, it is said to be autoregressive.

*Proceeding of "Technology Integration for Sustainable Development: An International E-Conference on Electrical, Electronics, Computer Science and Mechanical Engineering. (EECM-2023)". Organized by SJUIT.*

### 1.1 Machine Learning Models

### A. Long-Short Term Memory

A correct gradient-based learning method is used in conjunction with a revolutionary recurrent network design called LSTM. These issues with error back-flow are addressed by LSTM. Even in noisy cases and compact input sequences, it can learn to bridge time spans beyond 1000 steps without losing its short time delay skills. This is accomplished for an architecture requiring consistent error flow that is neither inflating nor disappearing through internal states of each unit using an effective, gradient-based technique.

In theory, an LSTMs can keep track of the different text qualities it is currently processing and recall distant information using the memory cells in its structure. For instance, writing gadget cell weights that would enable the cell to monitor if it is inside a quoted string is a straightforward exercise. The main objective of an organization's preparation cycle is to minimize misfortune (in terms of error or cost) experienced in the output when preparing information that is transmitted through it. It can calculate the likelihood, or misfortune, associated with a certain arrangement of loads, adjust those loads as necessary, and repeat the cycle until we find the perfect set of loads, for which misfortune is minimal.

Backtracking is the concept in question. Sometimes the slope turns out to be essentially irrelevant. It should be noted that a layer's slope depends on particular sections of the advancing layers. In the unlikely event that some of these segments are small (under 1), the acquired result is likely to be substantially more modest. The scaling impact is what is meant by this. This angle produces a moderate worth when combined with the learning rate, which has a small worth between 0.1 and 0.001. As a result, the change in loads is barely noticeable and the yield is almost identical to before. The phone was then enhanced by a few gating units and was called LSTM.

### B. Facebook Prophet Model

The prophet is a method for predicting time series data that uses additive modeling to fit non-linear patterns with seasonality that occurs annually, weekly, and daily. The data presented here should have good seasonal impacts and should include data from a number of seasons (or historical time periods). Prophet can handle missing data and trend changes.

### C. Auto Regressive Integrated Moving Average Model

The time-series auto-regressive technique known as Auto-Regressive Integrated Moving Average (ARIMA) determines short-term projections for the future by examining time-series of past data. For example, Hemorrhagic Fever with Renal Syndrome (HFRS), Hand, Foot, and Mouth Disease (HFMD), Hepatitis-B, and the recently discovered COVID-19 virus have all been predicted using ARIMA in the past.

Using several mathematical modeling approaches made prominent in epidemiology research, researchers have been examining the pattern and rate of COVID-19 infection since it first emerged in late 2019. Based on geolocation and data from the previous two weeks, an artificial neural network (ANN) prediction model was created to forecast the growth of COVID-19 instances globally. This study discovered that there was good agreement between the projected numbers generated by their model and the actual values. It is suggested to use the FPASSA-ANFIS model. The salp swarm algorithm (SSA) is used to augment the enhanced flower pollination algorithm (FPA) of the adaptive neuro-fuzzy inference system (ANFIS). When compared to several other current models using various accuracy measures, their model performed well.

## 2. Related Work

### 2.1 Coronavirus Disease (covid-19) cases Analysis using Machine-Learning Applications

Ameer Sardar et al. [1] states that the purpose of this study is to detect the role of machine-learning applications and algorithms in investigating and various purposes that deals with COVID-19. Their findings show that machinelearning can produce an important role in COVID-19 investigations, prediction, and discrimination.In conclusion, machine learning can be involved in the health provider programs and plans to assess and triage the COVID-19 cases. Supervised learning showed better results than other Unsupervised learning algorithms by having 92.9 percent testing accuracy. In the future recurrent supervised learning can be utilized for superior accuracy.

### 2.2 Optimizing Lstm For Time Series Prediction In Indian Stock Market

Anita Yadav et al.[2] states that Long Short Term Memory (LSTM)is one of the most popular deep learning models.

It is applied to time series prediction which is a particularly hard problem to solve due to the presence of long term trend, seasonal and cyclical fluctuations and random noise. The performance of LSTM was highly dependent on choice of several hyper-parameters which need to be chosen very carefully, in order to get good results. Being a relatively new model, there are no established guidelines for configuring LSTM. [2] In this paper this research gap was addressed. A datasetwas created from the Indian stock market and an LSTM model was developed for it. It was then optimized by comparing stateless and stateful models and bytuning for the number of hidden layers.

### 2.3 Maximum Powerdem And Prediction Using Fbprophet With Adaptive Kalman Filtering

C GUO et al. states that [10], It is very difficult to predict the Maximum Power Demand (MPD) of customers in high performance because ofvarious factors. In this paper, the problem of MPD prediction is studied by using fused machine learning algorithms. Firstly, an improved grey relation analysis method is adopted to analyze relevant influencing factors. Secondly, a modified prediction algorithm based on an adaptive cubature Kalman filter combined with Fbprophet is proposed according to the characteristics of customers' MPD. Finally, the proposed algorithm of this paper is applied to predict MPD and cost is evaluated. Experiment results show that the improved MPD prediction algorithm can comprehensively consider the relevant factors, and has good performance in time series prediction

## 4. Implementation

The suggested system's general system design is shown in Figure 4.1. The time series dataset is gathered and pre-processed before analysis. Following preprocessing, the data was divided into training and testing. The LSTM is used with the training dataset. The test dataset can be evaluated using the Prophet model, ARIMA, and Random Forest Regression model. Following the findings, the model may be utilized to predict next cases, and the findings can be evaluated using RMSE, MSE, and MAEmetrics.

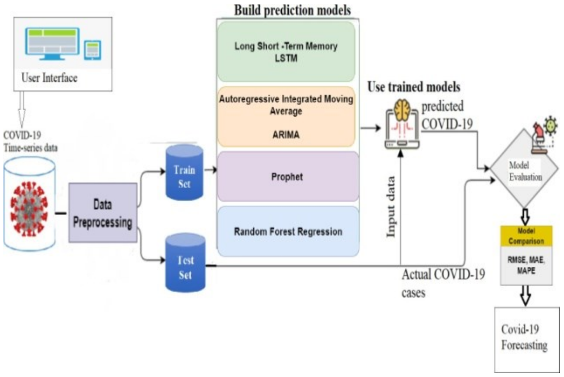


**Fig. 4.1 Model Evaluation**

It estimates how well (or how bad) the model is, in terms of its abilityin mapping the relationship between X (a feature, or independent variable,or predictor variable) and Y (the target, or dependent variable, or response variable).

### RMSE

The difference between values (sample or population values) predicted by a model or estimator and the values observed is typically measured using the root-mean-square deviation (RMSD) or root-mean-square error (RMSE).

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}\left(Predicted_i-Actual_i\right)^2}{N}}$$

**Equation 1. Root Mean squared error**

The RMSD is the quadratic mean of the square root of the second sample moment of the discrepancies between expected values and observed values.

When computations are made over the data sample that was used for estimate, these discrepancies are known as residuals. When calculated outside of a sample, these are known as mistakes (or prediction errors). The RMSD is used to combine the sizes of predictions' mistakes for different data points into a single indicator of predictive power. Since RMSD is scale-dependent, it should only be used to compare forecasting errors of several models for a single dataset and not between datasets.

A number of 0 (nearly never attained in practice) would represent a perfect fit to the data, and RMSD is always non-negative. A smaller RMSD is often preferable to a greater one. However, because the measure depends on the scale of the numbers used, comparisons across other types of data would be invalid.

The average of the squared errors' square root is the RMSD. Each error's impact on RMSD is proportionate to its size.

Therefore, RMSD is disproportionately impacted by bigger errors. RMSD is therefore vulnerable to outliers.

**MSE**

The average squared difference between the estimated values and true values is measured by an estimator's Mean Squared Error (MSE) or Mean Squared Deviation (MSD). It is a risk function that corresponds to the squared error loss's expected value. It is never negative, therefore numbers that are near to zero are preferable. The variance of the estimator and its bias are both incorporated into the MSE, which is the second moment of the error (around the origin).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

**Equation 2. Mean Squared Error**

**MAE**

A measure of mistakes between paired observations reflecting the same phenomena in statistics is called mean absolute error (MAE). Comparisons of predicted versus observed data, following time versus original data, and other examples of Y versus X.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

**Equation 3. Mean Absolute Error**

Time, as well as the comparison of one measurement method with another. The MAE is determined by dividing the total absolute errors by the sample size.

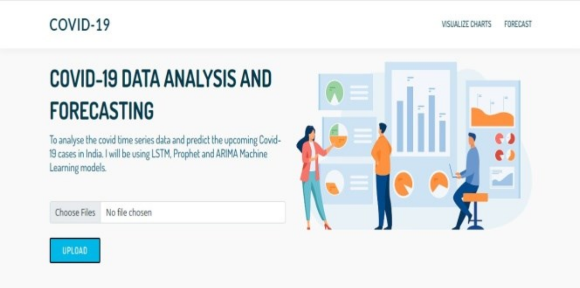
### Visualization and Forecasting

This system can forecast the COVID-19 instances for the following year by employing the aforementioned models and analyzing the assessment metrics. It can also visualize the graph of the results. For stunning results, our system made use of the matplotlib and plotly libraries.

## 5. Results and Analysis

### 5.1 Files Module

The shows the files module, which shows a list of uploaded files for analysis and forecasting.The files should only be in csv format, as this uses less memory and reads files more quickly.



COVID-19 DATA ANALYSIS AND FORECASTING

To analyse the covid time series data and predict the upcoming Covid-19 cases in India. I will be using LSTM, Prophet and ARIMA Machine Learning models.

Fig 5.1 Home Page

### 5.2 Machine Learning Models

The following machine learning models are used in this module to forecast the covid-19 situations. Each model will read the csv data, divide it into training and testing sets, and then train on the training data to make predictions about cases for the next days while also comparing the results to the testing dataset. The proposed system may obtain evaluation metrics and outcomes in MAE, MSE, and RMSE after comparison.

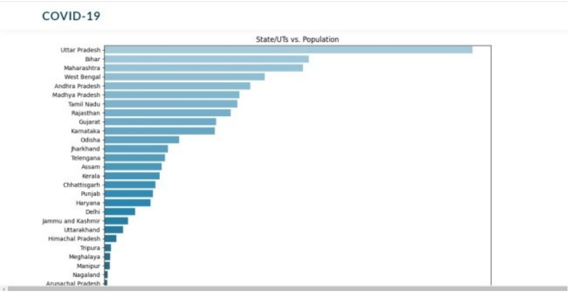
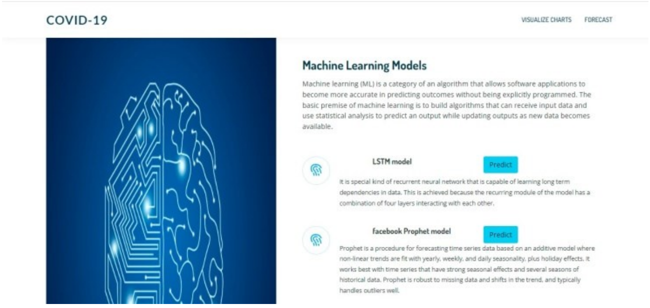


Fig 5.2 Data Files



Machine Learning Models

Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

LSTM model

It is special kind of recurrent neural network that is capable of learning long term dependencies in data. This is achieved because the recurring module of the model has a combination of four layers interacting with each other.

facebook Prophet model

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

Fig 5.3 Machine Learning Model

The Figure5.4 displays the total population of states inIndia.

### 5.3 Lstm Model Results

This module display how the LSTM model predicted the cases andhow it is performed in the process with comparison of actual cases.

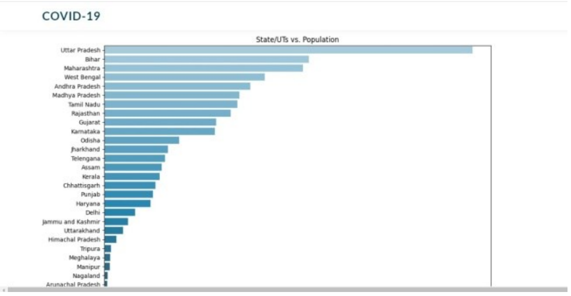


Fig.5.4 Population of UT and States

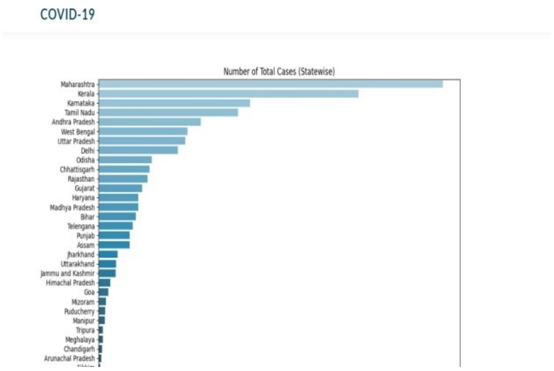
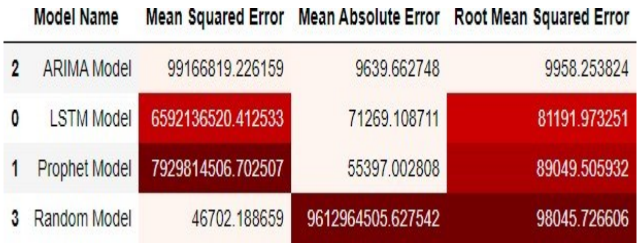


Fig 5.5 Maximum cases in India

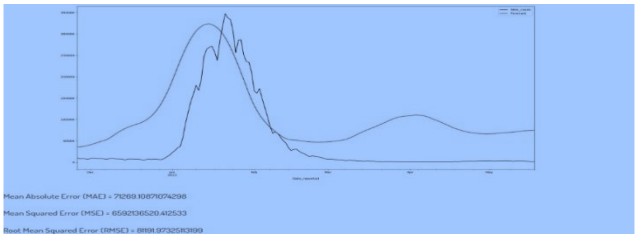


| | Model Name | Mean Squared Error | Mean Absolute Error | Root Mean Squared Error |
|---|---|---|---|---|
| 2 | ARIMA Model | 99166819.226159 | 9639.662748 | 9958.253824 |
| 0 | LSTM Model | 6592136520.412533 | 71269.108711 | 81191.973251 |
| 1 | Prophet Model | 7929814506.702507 | 55397.002808 | 89049.505932 |
| 3 | Random Model | 46702.188659 | 9612964505.627542 | 98045.726606 |

Fig 5.6 LSTM Results

### 5.4 Random Forest Model Results

This module display how the Random Forest model predicted thecases and how it is performed in the process with comparison of actual cases.



Mean Absolute Error (MAE) = 71269.10871074298

Mean Squared Error (MSE) = 6592136520.412533

Root Mean Squared Error (RMSE) = 81191.97325113199

Fig.5.6 RFM Results

### 5.5 Evaluation Metrics

The purpose of this module is to compare all the models and to choose the better model for future forecasting. The system can see in 4.19 that ARIMA model has performed better as it has low RMSE.

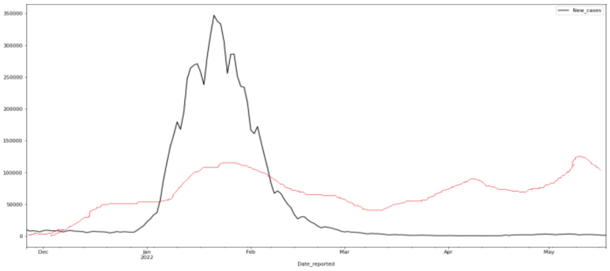


Fig. 5.7 Metrics on comparison

## Conclusion

The goal of this study was to utilize machine learning methods to forecast COVID-19 instances for the future year. The outcomes from several machine learning algorithms were able to forecast an increase in COVID-19 instances for the future year. In the COVID-19 examples, the ARIMA, LSTM, Prophet, and Random Forest all displayed promising outcomes. Using various models and algorithms, the suggested system's ML algorithms in health produced positive outcomes with high accuracy, sensitivity, and specificity. Prediction accuracy can be improved by using sources of data that are bivariate or multivariate. The data on deaths and recovered cases, the vaccination rate, and other COVID-19-related data can all contribute to a more accurate model of the disease trend in this instance.

## References

[1] Heamn N.Abduljabbar Ameer Sardar Kwekha Rashid and Bilal Alhayani. In *Coronavirus disease (COVID-19) cases analysis using machine-learning applications*, pages 501–513, 2020.

[2] C.K. AnitaYadav and Jha AditiSharan. In *Optimizing LSTM for time seriesprediction in Indian stock market*, pages 20–28, 2010.

[3] Tolga and Suleyman Serdar Kozat. Efficient online learning algorithms based on lstm neural networks. *IEEE transactions on neural networks andlearnings Systems Journal*, 29(8):3772–3783, 2017.

[4] Sweeti Sah B. Surendiran R. Dhanalakshmi Sachi Nandan Mohanty Fayadh Alenezi and Kemal Polat. In *Forecasting COVID-19 Pandemic Using Prophet, ARIMA, and Hybrid Stacked LSTM-GRU Models in India*,pages 78–92, 2022.

[5] Mohan Mahanty K. Swathi K. Sasi Teja P. Hemanth Kumar and A. Sravani. In *Forecasting the Spread of COVID-19 Pandemic with Prophet*, pages 115–122, 2021.

[6] Naresh Kumar and Seba Susan. Covid-19 pandemic prediction using timeseries forecasting models. *11th ICCCNT 2020 conference*, 25(5):159–187,2020.

[7] Sachin Aryal Ishan Manandhar Patricia B. MunroeAmeer Sardar Kwekha Rashid Ahmad Alimadadi. Artificial intelligence and machine learning to fight covid-19. In *AI and Machine Learning for Understanding Biological Processes*, pages 51–73, 2020.

[8] S.M Liu H. Li and C.K Tang. Coronavirus disease 2019 (covid-19): current status and future perspectives. *International Journal of Antimicrobial Agents*, 29(8):257–289, 2020.

[9] Shaun S Wulff. Time series analysis: Forecasting and control. In *Journal of Quality Technology*, 2017.

[10] C. Guo H. Jiang H. Yao. Maximum power demand prediction using fbprophet with adaptive kalman filtering. *IEEE Access*, 29(8):19236–19247, 2020.

## Author Biography

Ms.P.S.Apirajitha is working as a Assistant Professor in sri Venkateswara college of Engineering and has 16 years of Teaching Experience, Expertise in Cloud Computing, Blockchain Technologies and Security. She is currently pursuing Ph.D degree, and interested to learn new things.

Ms.G.Mahalakshmi is working as Teaching Professor in College of Engineering, Anna University, Chennai, has 14 years of Teaching Experience and expertise in the field of Networking, NLP and Data Mining. She is currently pursuing Ph.D degree.

Mr.M.Duraimurugan is working as Teaching Professor in College of Engineering, Anna University, Chennai, has 15 years of Teaching Experience and expertise in the field of Networking, and Data Mining. he is currently pursuing Ph.D degree